The Relevance of Self-Locating Beliefs

Michael G. Titelbaum
University of Wisconsin–Madison

How much do I learn when I learn what time it is, or where I am, or who I am? Beliefs about one's spatiotemporal location and beliefs about one's identity are often called "self-locating" beliefs.[1] Typically when an agent learns self-locating information she learns some non-self-locating information as well. Consider an agent who is on a long plane flight and isn't sure when the plane will land. If she learns that it is now 6 p.m., she also learns that the flight lands after 6 p.m. One might argue that in this example the self-locating information is learned and then the non-self-locating information is *inferred*, but for simplicity's sake we will count both information that is explicitly presented to an agent and what she can infer from that information given her background as "learned."[2] Even after we lump all this information together as learned, there are non-self-locating conclusions beyond what the agent learns in which she

I would like to thank Johan van Benthem, Darren Bradley, Fabrizio Cariani, David Chalmers, Mark Colyvan, Kenny Easwaran, Adam Elga, Justin Fisher, Branden Fitelson, Hilary Greaves, Alan Hájek, Joseph Halpern, Carrie Jenkins, Peter Lewis, John MacFarlane, Christopher Meacham, Sarah Moss, Jim Pryor, Sherrilyn Roush, Brian Skyrms, Susan Vineberg, Peter Vranas, two anonymous referees for the *Philosophical Review*, and audiences at the 2007 meeting of the Society for Exact Philosophy, the 2006 Australasian Association of Philosophy conference, the 2006 Formal Epistemology Workshop, the 2006 Paris-Berkeley-Amsterdam Logic Meeting, and the 2005 Berkeley-Stanford-Davis Graduate Student Conference.

1. David Lewis (1979) proposed thinking of all beliefs as self-locating: some locate the agent within the space of possible worlds, while others spatiotemporally locate the agent within a given possible world. Contemporary authors tend to refer to the latter as "self-locating" beliefs but not the former; I will follow that usage here.

2. We will also overlook the factivity of "learns." This will make no difference to our discussion because in all the stories we consider the information gained by agents is true.

Philosophical Review, Vol. 117, No. 4, 2008
DOI: 10.1215/00318108-2008-016

might rationally change her *degree* of belief. For example, when our flyer learns it is 6 p.m., she might increase her degree of belief that dinner is served on this flight.

Can an agent learn self-locating information without learning anything non-self-locating? Imagine a situation in which an agent is certain of every experience she's going to have between two times. She's watching a movie, say, that she's seen a hundred times and has thoroughly memorized; she's at home alone on her well-worn couch in the dark; the phone is unplugged; et cetera. It's not quite right to say that the agent learns *nothing* as the movie progresses—after all, she learns that some time has passed. But all she learns is self-locating information; she has no basis for inferring new non-self-locating conclusions. We may have the intuition that she should keep her non-self-locating *degrees* of belief constant as well. Described in possible world terms, the intuition is this: since the agent's information concerns only her movement from one spatiotemporal location to another *within* a possible world, she should not change her degrees of belief *across* possible worlds.

As we will see, a number of philosophers have had this intuition. Generalized, it supports the following thesis:

> **Relevance-Limiting Thesis**: It is never rational for an agent who learns only self-locating information to respond by altering a non-self-locating degree of belief.

According to the Relevance-Limiting Thesis, if a piece of self-locating information cannot rule out any non-self-locating possibilities entertained by an agent, it should not have more subtle effects on her non-self-locating degrees of belief either.

This essay argues that the Relevance-Limiting Thesis is false. My strategy is to provide a counterexample: a story in which the agent clearly learns nothing non-self-locating between two times and yet in which she is rationally required to respond to the self-locating information she learns by altering a non-self-locating degree of belief.

Executing this strategy is difficult not because the story is novel (it has been discussed in the decision theory and Bayesian literature for years), but because we lack the tools to settle what rationality requires of the agent involved. For some time formal epistemologists have possessed precise principles for modeling rational degree-of-belief changes in response to new evidence. Unfortunately, these principles systematically fail when applied to stories involving self-locating beliefs. This is

because self-locating beliefs are context-sensitive, and traditional belief-update rules were not designed to handle context-sensitivity.

In this essay I present a new formal framework for modeling rational degree-of-belief updates over time. Unlike traditional Bayesian techniques, this framework can accurately model changing rational degrees of belief in situations involving context-sensitivity. I begin by describing the structure of my models and some well-established principles for modeling synchronically rational agents. I then discuss the most popular Bayesian principle for modeling diachronically rational agents, updating by conditionalization, and explain why it yields inaccurate verdicts for stories involving context-sensitive beliefs. To correctly model these stories, I introduce two new principles: a restricted updating principle and a principle relating multiple models of the same story with different modeling languages. Once the new principles are described, I show that they give intuitively correct results for some simple stories, I suggest more abstract reasons why they deserve our confidence, and I explain how to use them to analyze stories of various types.

Finally I present a counterexample to the Relevance-Limiting Thesis, a story known as the Sleeping Beauty Problem. I show that if we apply my framework, the controversial part of the problem can be resolved without appeal to frequencies, objective chances, indifference principles, utilities, or Dutch Books. A correct framework for modeling context-sensitive beliefs is sufficient by itself to demonstrate that upon awakening and learning some purely self-locating information, Beauty should decrease her non-self-locating degree of belief that a particular coin flip came up heads. This in turn is sufficient to disprove the Relevance-Limiting Thesis.

1. The Modeling Framework

1.1. Basic Structure

Our modeling framework is designed to model what I call "stories." A **story** describes an agent who starts off with a particular set of certainties and then becomes certain of other things at various times. The modeling framework aims to determine what rational requirements these evolving sets of certainties impose on the agent's nonextreme degrees of belief. While our stories will often involve agents' assigning certainty to empirical propositions, this is not because I think it is rationally permissible for real-world agents to be absolutely certain of matters empirical. Rather, stipulated certainties simplify doxastic problems for formal

analysis. To allow us to focus on the degrees of belief most at issue, we imagine that the agent assumes various other things to be true. These assumptions appear in our stories as certainties.[3]

The core idea of Bayesian modeling is to represent an agent's degrees of belief with the values of a real-valued function called a "credence function." For a story in which the agent's degrees of belief evolve over time, we will need separate credence functions for each time of interest during the story. So to model a story, we start by selecting a time set. The **time set** of a model is a nonempty, finite set of moments $\{t_1, t_2, \ldots, t_n\}$ during the story (with subscripts reflecting their temporal order) at which we will model the agent's degrees of belief.

Next, we specify a modeling language over which our credence functions will be defined. Credence functions have traditionally been defined over sets of propositions. But if "Bruce Wayne lives in Wayne Manor" and "Batman lives in Wayne Manor" express the same proposition, credence functions defined on propositions will be unable to model an agent who assigns them different degrees of belief at the same time. Similarly, "It is sunny today" and "It is sunny on Monday" may express the same proposition on Monday, but we want to be able to model an agent who assigns them different degrees of belief on that day. Thus our credence functions will be defined over sets of "sentences." A **sentence** is a syntactical object consisting of a string of symbols. Since it can be rational for an agent to assign different degrees of belief to "Bruce Wayne lives in Wayne Manor" and "Batman lives in Wayne Manor" at the same time, they will be represented in our modeling language by distinct sentences. "It is sunny today" and "It is sunny on Monday" will also be represented by distinct sentences, but we will not have distinct sentences representing "It is sunny today" on Monday and "It is sunny today" on Tuesday. This is unnecessary because our framework can assign different credences to the same sentence at different times.

Credences in sentences ultimately represent doxastic states of the agent. However, our framework need not take a position on whether the objects of doxastic states are propositions (centered or uncentered), propositions-under-descriptions, linguistic entities, or something else. So I will describe sentences in the modeling language as representing

3. Mark Lance (1995) argues that a Bayesian model—indeed, *any* type of explicit decision-theoretic model—of a situation must always work within a structure of empirical propositions that the agent is assumed to accept prior to the application of the model.

"claims." A **claim** is a natural-language sentence that takes a truth-value in a context, such as "It is sunny today." When I discuss an agent's "degree of belief in 'It is sunny today'," or her "degree of belief that it is sunny today," these locutions refer to her degree of belief that the claim "It is sunny today" is true in the current context. We might elicit this degree of belief by asking the agent something like "How confident are you right now that it is sunny today?" The philosophy of mind and language can then provide a deeper account of what kind of thing the agent's answer is *about*.

To specify a modeling language, we specify a nonempty finite set of atomic sentences and the claims they represent. From the point of view of the model, **atomic sentences** are primitives with no internal structure. The **modeling language** is the set containing the atomic sentences and any sentences that can be formed from them by applying symbols for truth-functions in the standard iterative fashion. The truth-functional symbols in the sentences of a modeling language represent truth-functional connectives within claims: for example, a conjunctive sentence in the modeling language represents the claim that is the conjunction of the claims represented by the sentence's conjuncts. Since truth-functional structure is the only logical structure represented in the modeling language, metatheoretical statements we make about logical relations (equivalence, entailment, mutual exclusivity, etc.) between *sentences* should be read as concerning syntactical relations of classical propositional logic only. (Two sentences may therefore be described as "nonequivalent" even though they represent claims that are equivalent in, say, first-order logic.) Note that every modeling language will contain at least one tautology and one contradiction.

Our models will involve two kinds of credence functions. An **unconditional credence function** is a function from sentences in the modeling language to the reals. A value of an unconditional credence function represents the agent's degree of belief at a given time in a particular claim. For example, $P_1(x)$ represents the agent's degree of belief at time t_1 in the claim represented by x. A higher credence value represents greater confidence in a particular claim, and an unconditional credence of 1 represents certainty of that claim. A **conditional credence function** is a partial function from ordered pairs of sentences in the modeling language to the reals. The conditional credence $P_1(x \mid y)$ represents the agent's degree of belief at time t_1 in the claim represented by x conditional on the supposition of the claim represented by y. A **history** is a set of credence functions, containing exactly one conditional

credence function and one unconditional credence function indexed to each time in the time set.

Credence functions are subject to two types of constraints. **Systematic constraints** are common to every model we build using this modeling framework, regardless of what story that model represents. I view these constraints, taken together, as representing consistency requirements of rationality. Developing our modeling framework's systematic constraints is the main task of this essay's first half.

Rational requirements not represented by the systematic constraints are represented by **extrasystematic constraints**. Some extrasystematic constraints compensate for the limitations of the propositional logic underlying the systematic constraints: for example, the rational requirement that an agent be certain of "I am here now" will be implemented by an extrasystematic constraint.[4] But most extrasystematic constraints represent rational requirements derived from the specific details of the story being modeled. For example, if in a story the agent is certain at t_1 that a particular coin flip is fair but has no evidence of its outcome, David Lewis's Principal Principle (1980) suggests that rationality requires the agent's degree of belief that the coin comes up heads to be 1/2. This requirement would be represented with the extrasystematic constraint $P_1(h) = 1/2$.

Among the extrasystematic constraints on a model are constraints representing the evolving sets of certainties stipulated by the story. For each sentence x in the modeling language and each time t_k in the time set, there will either be an extrasystematic constraint that $P_k(x) = 1$ or an extrasystematic constraint that $P_k(x) < 1$. The extrasystematic constraints will assign x a credence of 1 at t_k just in case the story stipulates that the agent is certain at t_k of the claim represented by x, or x represents a claim entailed by claims stipulated as certain at t_k.[5] The claims doing the entailing need not be claims represented by sentences in the modeling language, and the entailments need not be truth-functional entailments (they can involve quantifiers, the logic of indexicals, etc.). Thus a sentence representing a logical truth of any kind will always receive a credence of 1.

4. This approach follows Garber 1983.

5. Such stipulations will sometimes be implicit in the story's context. For example, in a story not involving Cartesian-style skepticism we will take "I am now awake" to be stipulated as certain at all times.

A model of a story is defined relative to a particular time set, modeling language, and set of extrasystematic constraints. The **model** is the set of all possible histories (with credence functions defined relative to that time set and that modeling language) that meet both the systematic and extrasystematic constraints.

We will often describe features of a model using arithmetic statements. An **arithmetic statement** for a model is an equality or inequality relating two expressions composed arithmetically from credence values in that model and/or constants. An arithmetic statement contains no variables or quantifiers. For example, if x and y are sentences in a model's modeling language and t_1 and t_2 are times in its time set, $P_1(x) + P_2(x \mid y) = 1/2$ is an arithmetic statement for that model. Since conditional credence functions may be partial functions, we will also allow arithmetic statements of the form "$P_1(x \mid y)$ is undefined."

An arithmetic statement for a model that is true in every history of that model is called a **verdict** of that model. The extrasystematic constraints on a model are verdicts of that model; instances of our systematic constraints will also be verdicts. All verdicts of a model are either extrasystematic constraints or instances of systematic constraints or can be derived algebraically from verdicts of those two kinds.

A model evaluates the evolving doxastic states of an agent through its verdicts. If the arithmetic statement for a model representing some feature of the agent's evolving doxastic state contradicts a verdict of that model, the model evaluates the evolution of that agent's doxastic state as violating a requirement of ideal rationality. Thus a model's verdicts represent what it takes to be necessary conditions for ideal rationality. For example, suppose model M yields the verdict $P_0(c) < 0.5$. If the agent in the story assigns a degree of belief of 0.7 at t_0 to the claim represented by c, M represents that agent's doxastic state as violating a requirement of ideal rationality. Also, if M yields the verdict that $P_1(a \mid b)$ is undefined (that is, if that conditional credence is undefined in every history of M), then if at t_1 the agent assigns a conditional degree of belief to the claim represented by a on the supposition of the claim represented by b, M represents her state as violating a requirement of ideal rationality.

A model's verdicts represent *necessary* conditions for an evaluative standard I call "ideal rationality." (The standards of ideal rationality are stronger than the standards we require for rationality in everyday conversation; for example, an agent's doxastic state violates the requirements of ideal rationality if she is certain of two logically contradictory

claims, even though in everyday parlance we might not evaluate her state as irrational if she is unaware that they are contradictory.) These verdicts will sometimes be strong enough to limit the agent to exactly one permissible degree of belief assignment to a claim. When the verdicts are not that strong, we will interpret the various histories contained in the model as distinct possible evolutions of the agent's doxastic state, each of which the model deems in compliance with the requirements of ideal rationality. For example, if model M yields the verdict $P_0(c) < 0.5$, by M's lights it is ideally rational for the agent to assign any degree of belief less than 0.5 at t_0 to the claim represented by c. Alternatively, we might interpret such a verdict as representing a requirement of ideal rationality that the agent adopt a doxastic state toward c represented by the full *range* of reals from 0 to 0.5. But working with such ranged doxastic states, sometimes described as "imprecise probability assignments," requires a more intricate apparatus for interpreting verdicts than I want to develop here. So we will set aside ranged doxastic states for the rest of this essay.

Finally, a word on notation: While I will rely on context to distinguish use from mention in most cases, I will use notation to distinguish the various elements of our modeling framework. An uppercase unitalicized character or a string of such characters will name a model (for example, "model M"). A string of italicized characters will be an atomic sentence in a modeling language ("*MonRed*"). A single uppercase italicized letter will typically name a set of sentences ("L"). The exceptions to this rule are "P," which represents credence functions, and "T," which represents time sets. When discussing multiple models at once, I will differentiate them using superscripts ("model M^+"). A model's modeling language will always be named with an "L" followed by the model name's superscript, its credence functions with a "P" then the superscript, and its time set with a "T" followed by the superscript. Finally, I will **bold** a technical term when defining it.

1.2. Synchronic Constraints

The first four systematic constraints on our models are "synchronic" constraints. Taken together, they represent consistency requirements on sets of degrees of belief assigned at the same time. Systematic constraints (1) through (3), Kolmogorov's axioms, require each unconditional $P_k(\cdot)$ to be a **probability function**. For a model M, these constraints are:

Systematic Constraints (1)–(3), Kolmogorov Axioms:

(1) For any $t_k \in T$ and any sentence $x \in L$, $P_k(x) \geq 0$.

(2) For any $t_k \in T$ and any tautological sentence $\mathsf{T} \in L$, $P_k(\mathsf{T}) = 1$.

(3) For any $t_k \in T$ and any mutually exclusive sentences $x, y \in L$,

$$P_k(x \vee y) = P_k(x) + P_k(y).$$

These axioms restrict all credence values to the range $[0, 1]$. With the axioms in place, a credence of 0 in x represents certainty in the claim represented by $\sim x$.

The next synchronic constraint relates conditional credences to unconditional credences indexed to the same time. For a model M, the constraint is:

> **Systematic Constraint (4)**:
> For any $x, y \in L$ and any $t_k \in T$, if $P_k(y) > 0$, then $P_k(x \mid y) = \frac{P_k(x \,\&\, y)}{P_k(y)}$. If $P_k(y) = 0$, $P_k(x \mid y)$ is undefined.

The $P_k(y) = 0$ clause in this constraint explains the role of undefined conditional credences in our framework: if an agent assigns a degree of belief of zero to a claim, she violates a requirement of ideal rationality if she assigns a degree of belief to any claim conditional on that claim.[6]

1.3. Conditionalization

Our next task is to identify a "diachronic" systematic constraint on our models. By combining the diachronic constraint with our synchronic constraints, we can represent consistency requirements on sets of degrees of belief assigned at different times. The best-known diachronic constraint on degrees of belief is a principle called "updating by conditionalization," which is usually stated something like:

> **Conditionalization (preliminary)**: An agent's credence in x at t_k is her credence in x at an earlier time t_j conditional on everything she learns between t_j and t_k.

6. One might object to systematic constraint (4) by appealing to infinitistic stories in which a credence of 0 in x can fail to represent certainty that the claim represented by x is false. (This is one of the objections to systematic constraint (4) made in Hájek 2003.) Our modeling framework is not designed to apply to infinitistic stories, and so assumes that a credence of 0 in x represents certainty in the claim represented by $\sim x$.

To apply this constraint in a formal setting, we need to formulate it more precisely. This will require a bit more notation. First, given model M and time $t_i \in T$, the agent's **certainty set** at t_i is defined as the set $C_i = \{x \in L : P_i(x) = 1\}$. (For model M^+, the certainty set at t_i will be written C_i^+.) Given two times t_j and t_k in T with $t_j < t_k$, I will sometimes refer to the set $C_k - C_j$ as the **gained certainty set** between t_j and t_k. Second, the brackets "⟨" and "⟩" will indicate a function from subsets of L to elements of L. If $S \subseteq L$ is nonempty, $\langle S \rangle$ is a sentence in L logically equivalent to the conjunction of the sentences in S. If S is empty, $\langle S \rangle$ is a tautology in L.[7]

With this notation in place, we can formulate Conditionalization as:

> **Conditionalization**: Given a model M, any $t_j, t_k \in T$ with $j < k$, and any $x \in L$, $P_k(x) = P_j(x \mid \langle C_k - C_j \rangle)$.

This formulation gives the quantifier in "everything she learns" a precise domain: the modeling language of the model with which we are working. This puts pressure on us to choose an appropriate modeling language; if the modeling language fails to represent relevant claims in the story, Conditionalization may yield verdicts that do not represent requirements of ideal rationality. This issue will be discussed in detail in section 1.5.

The precise formulation of Conditionalization yields intuitive verdicts for a variety of stories. For example, consider this story:

> *The Die:* Marilynn walks into a room. She is told that a few minutes ago a fair die was thrown, and in a few minutes a loudspeaker will announce whether the outcome was odd or even. A few minutes later, the loudspeaker announces that the die came up odd. Assuming Marilynn believes everything she has heard with certainty, what does ideal rationality require of the relationship between her post-announcement degree of belief that the die came up 3 and her preannouncement degrees of belief?

7. For any S there will be multiple elements of L meeting this description. Given our synchronic constraints, it will not matter which serves as $\langle S \rangle$.

Table 1. Model D

Story: The Die	ES: (1) $0 < P_1(\mathit{Three}) < 1$
	(2) $0 < P_2(\mathit{Three}) < 1$
T: Contains these times:	(3) $0 < P_1(\mathit{Odd}) < 1$
t_1 After Marilynn is told about the die but before she hears the announcement.	(4) $P_2(\mathit{Odd}) = 1$
	(5) $P_1(\mathit{Three} \supset \mathit{Odd}) = 1$
t_2 After Marilynn hears the announcement.	GCS: $\langle C_2 - C_1 \rangle \dashv\vdash \mathit{Odd}$
L: Built on these atomic sentences, representing these claims:	
Three The die came up 3.	
Odd The die came up odd.	

The most straightforward model for this story, model D, is described in table 1. Note that the list of extrasystematic constraints there ("ES") is not exhaustive; in describing models I will list only the most pertinent extrasystematic constraints. For efficiency's sake I have also used our synchronic systematic constraints to represent multiple extrasystematic constraints in one "double inequality"; for example I have combined $P_1(\mathit{Three}) < 1$ and $P_1(\sim\mathit{Three}) < 1$ into $0 < P_1(\mathit{Three}) < 1$. Finally, while it is redundant in describing the model (because it follows from a full list of extrasystematic constraints), for the sake of convenience I have indicated the gained certainty set ("GCS") between t_1 and t_2 by specifying a sentence syntactically equivalent to $\langle C_2 - C_1 \rangle$.

If we adopted Conditionalization as a systematic constraint of our modeling framework, model D would yield the following verdict:

$$P_2(\mathit{Three}) = P_1(\mathit{Three} \mid \mathit{Odd}). \tag{1}$$

Intuitively, this verdict represents a requirement of ideal rationality. We can combine equation (1) with our synchronic constraints and extrasystematic constraints to derive

$$P_2(\mathit{Three}) > P_1(\mathit{Three}), \tag{2}$$

which captures the rational response to learning that the die came up odd.

We could get even more precise information out of equation (1) by applying the Principal Principle. Since Marilynn is certain at t_1 that the die is fair, we could add to the model the extrasystematic constraints

$P_1(Odd) = \frac{1}{2}$ and $P_1(Three) = \frac{1}{6}$. Together with our other constraints, these would yield

$$P_2(Three) = P_1(Three \mid Odd) = \frac{P_1(Three \,\&\, Odd)}{P_1(Odd)}$$
$$= \frac{P_1(Three)}{P_1(Odd)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}, \tag{3}$$

which also squares with our intuitions about the case.

1.4. Limited Conditionalization

While Conditionalization yields intuitive verdicts for a variety of stories, it can fail when applied to stories involving context-sensitive claims. For example, consider the following story:

> *Sleeping In:* Olga awakens one morning to find her clock blinking "6 a.m." The blinking indicates that some time during the night the clock reset, but Olga is certain it isn't more than a few hours off. She goes back to sleep, and when she awakens again the clock blinks "2 p.m." How should Olga's degree of belief that it is now afternoon relate to the degrees of belief she had on her first awakening?

A model for this story, model SI, is described in table 2.

Sleeping In asks us to relate $P_2(pm)$ to P_1 values. If Conditionalization were a systematic constraint of our modeling framework, it would yield the following verdict:

$$P_2(pm) = P_1(pm \mid Two \,\&\, 2^{nd}\,Two \,\&\, [pm \equiv 2^{nd}\,pm]). \tag{4}$$

But this verdict does not represent a requirement of ideal rationality. It links Olga's t_2 degree of belief that it is afternoon at that moment to her conditional t_1 degree of belief that it is afternoon at *that* moment in a way that doesn't make sense. Even worse, since $P_1(Two) = 0$, our synchronic systematic constraints will give the sentence conditioned on in equation (4) an unconditional t_1 credence of 0 and therefore make $P_2(pm)$ undefined. In fact, we could generate a similar verdict requiring *any* of Olga's t_2 degrees of belief to be undefined!

Conditionalization was designed to model situations in which a degree of belief that goes to one extreme stays at that extreme. But in

Table 2. Model SI

Story: Sleeping In	$2^{nd}Two$ The second time Olga awakens, the clock reads 2 p.m.
T: Contains these times:	$2^{nd}pm$ The second time Olga awakens, it is afternoon.
t_1 After Olga first awakens and sees the clock reading 6 a.m.	ES: (1) $P_1(Two) = 0$
t_2 After Olga awakens for the second time and sees the clock reading 2 p.m.	(2) $P_2(Two) = 1$
	(3) $P_1(pm) = 0$
	(4) $0 < P_2(pm) < 1$
	(5) $P_1(1^{st}Two) = 0$
L: Built on these atomic sentences, representing these claims:	(6) $P_1(1^{st}pm) = 0$
	(7) $0 < P_1(2^{nd}Two) < 1$
Two The clock now reads 2 p.m.	(8) $P_2(2^{nd}Two) = 1$
pm It is now afternoon.	(9) $0 < P_1(pm \equiv 2^{nd}pm) < 1$
$1^{st}Two$ The first time Olga awakens, the clock reads 2 p.m.	(10) $P_2(pm \equiv 2^{nd}pm) = 1$
$1^{st}pm$ The first time Olga awakens, it is afternoon.	GCS: $\langle C_2 - C_1 \rangle \dashv\vdash Two \,\&\, 2^{nd}Two \,\&\, (pm \equiv 2^{nd}pm)$

stories involving context-sensitive claims, it can be rational for an agent's degree of belief in a claim to go from 1 to something less, or from 0 to something greater. The relevant feature of context-sensitive claims is their ability to have different truth-values in different contexts. Because of these shifting truth-values, an agent who assigns a degree of belief of 1 to "Today is Monday" and 0 to "Today is Tuesday" at a particular time can reverse those assignments twenty-four hours later without violating requirements of ideal rationality.

In Sleeping In, Olga becomes certain at t_2 of a claim ("The clock now reads 2 p.m.") she was certain at t_1 was false. This is rational because Olga is aware that this claim is context-sensitive. Yet because the sentence representing this claim had an unconditional P_1 value of 0, applying Conditionalization between t_1 and t_2 yields verdicts like equation (4) that require Olga's t_2 credences to be undefined.

One might object that I have caused this problem myself, either by focusing on Conditionalization as a diachronic constraint or by using a modeling language that represents context-sensitive claims with the same sentence at different times. Yet other popular diachronic constraints, such as Jeffrey Conditionalization (Jeffrey 1983) and the

Reflection Principle (van Fraassen 1995), also run into problems when applied to stories involving context-sensitive claims. We might avoid some of Conditionalization's troubles by representing the same claim with different sentences at different times in the time set, but then Conditionalization would become nearly useless in relating credences at one time to credences at the next. The resulting framework would need much more machinery to yield any significant verdicts.

The simplest response to Conditionalization's difficulties with context-sensitive claims is to limit Conditionalization, restricting it from applying to cases we know will give it trouble. We can do so with the following diachronic constraint, which will be our framework's fifth systematic constraint:

> **Systematic Constraint (5), Limited Conditionalization (LC)**:
> Given a model M, any $t_j, t_k \in T$ with $j < k$, and any $x \in L$, if $C_j \subseteq C_k$, then $P_k(x) = P_j(x \mid \langle C_k - C_j \rangle)$.

(LC) is Conditionalization with one added condition: (LC) relates credences at two times only when the earlier time's certainty set is a subset of the later's.[8] Why adopt this condition in particular? On the one hand, it might seem like an overreaction. The trouble in Sleeping In was caused by sentences' going from one extremal credence to the other. But (LC) declines to relate credences at two times when a sentence goes from a credence of 1 to *any* lower credence, even if that lower credence is not 0. If we adopted a less-limited conditionalization principle—one that conditionalized in all cases except those in which a credence went from one extreme to the other—our framework would give us more verdicts about the stories we analyze. On the other hand, (LC) might be an *under*reaction: even with (LC)'s strong antecedent, there still might be cases for which the constraint yields verdicts that do not represent requirements of ideal rationality.

Those two concerns will be addressed below, in sections 1.7 and 1.8, respectively. For now, notice that in the stories we've considered so far (LC) recovers verdicts we want without generating any we don't. (LC) easily yields the intuitive verdict for The Die that we derived in equation (1) because in that story $C_1 \subseteq C_2$. But in model SI, C_1 is not a

8. M. J. Schervish, T. Seidenfeld, and J. Kadane (2004) note that a similar limit on conditionalizing has been accepted in the statistics literature for some time.

subset of C_2. (For example, $\sim Two \in C_1$ but $\sim Two \notin C_2$.) Thus (LC) does not place any constraints on the relation between P_1 and P_2, and so does not generate the absurd equation (4). Unlike Conditionalization, (LC) will not generate verdicts for Sleeping In that fail to represent requirements of ideal rationality.

Yet there is clearly more work to do. (LC) avoids generating incorrect verdicts for SI by not generating any diachronic verdicts for SI at all. But surely in the Sleeping In story there are rational constraints on the relations between Olga's t_1 and t_2 degrees of belief. To represent those relations in our model, we will need our sixth and final systematic constraint.

1.5. Proper Expansion Principle

Our precise formulation of Conditionalization in section 1.3 relativized that constraint to the particular modeling language used in a model. (LC) is also relativized to modeling languages in this way. This raises the threat that by choosing a modeling language that fails to represent all the relevant claims in a story we might construct a model whose verdicts do not represent requirements of ideal rationality.

No modeling language can represent every possible claim. Thus our best option is to select a modeling language that represents all the relevant claims we can think of, but be open to the suggestion that a particular omitted claim is also relevant. If such a suggestion is made, we respond by extending our modeling language to include a sentence representing the putatively relevant claim. We then check to see if the verdicts of the resulting model replicate those of the original.

The modeling framework presented here facilitates such conversations about choice of modeling language. Bayesians often leave decisions about which claims to include implicit in discussing their models. But our framework requires an explicit statement of the modeling language in the course of specifying a model. The choice of modeling language becomes a parameter of the model that we can debate and adjust accordingly.

Our final systematic constraint concerns models of the same story with different modeling languages. The constraint keeps a model's verdicts intact when its modeling language is extended in particular ways. To introduce that constraint, we need more terminology.

An arithmetic statement for one model has an **analogue** for another model obtained by replacing the statement's superscripts with the superscripts of the other model (or, if there are none, by adding in

the relevant superscripts). For example, if $P_1(x) = 1/2$ is an arithmetic statement for model M its analogue for M^+ is $P_1^+(x) = 1/2$. Informally, an arithmetic statement in one model and its analogue in another "say" the same thing about the agent's degrees of belief.

Given two models M and M^+, we call M^+ an **expansion** of M just in case

- $L \subseteq L^+$,
- $T = T^+$, and
- An arithmetic statement for M is an extrasystematic constraint on M just in case its analogue for M^+ is an extrasystematic constraint on M^+.

If M^+ is an expansion of M, we call M a **reduction** of M^+.

Now suppose M^+ is an expansion of M. M^+ is a **proper expansion** of M just in case for any $y \in L^+$ and any $t_k \in T^+$, there exists an $x \in L$ such that $P_k^+(x \equiv y) = 1$. To be clear about the quantifier order, M^+ is a proper expansion of M just in case M^+ is an expansion of M and

$$(\forall y \in L^+)(\forall t_k \in T^+)(\exists x \in L)(P_k^+(x \equiv y) = 1).$$

If M^+ is a proper expansion of M, we will call M a **proper reduction** of M^+.

Our final systematic constraint creates conditions under which extending a model's modeling language keeps its verdicts intact.

> **Systematic Constraint (6), Proper Expansion Principle (PEP)**:
> If M^+ is a proper expansion of M, the analogue for M^+ of any verdict of M is a verdict of M^+.

The rest of this section, along with the entirety of the next two, is devoted to showing that (PEP) yields verdicts that represent requirements of ideal rationality. We'll start by returning to our story The Die.

Our original model of The Die (model D) had a modeling language containing two atomic sentences (*Three* and *Odd*) representing claims about the outcome of the die roll. Yet between t_1 and t_2 Marilynn becomes certain not only that the die came up odd, but also that some time has passed since the situation was explained to her. For example, she becomes certain of the claim "The odd/even announcement has been made." It might be objected (by someone particularly attentive to self-locating beliefs) that this information about the passage of time is relevant to the other claims represented in D, and that by failing to

Table 3. Model D^+

Story: The Die	*Made* The odd/even announcement has been made.
T^+: Contains these times:	
t_1 After Marilynn is told about the die but before she hears the announcement.	ES: (1) $0 < P_1^+(\mathit{Three}) < 1$
	(2) $0 < P_2^+(\mathit{Three}) < 1$
	(3) $0 < P_1^+(\mathit{Odd}) < 1$
t_2 After Marilynn hears the announcement.	(4) $P_2^+(\mathit{Odd}) = 1$
	(5) $P_1^+(\mathit{Three} \supset \mathit{Odd}) = 1$
	(6) $P_1^+(\mathit{Made}) = 0$
L^+: Built on these atomic sentences, representing these claims:	(7) $P_2^+(\mathit{Made}) = 1$
Three The die came up 3.	GCS: $\langle C_2^+ - C_1^+ \rangle \dashv\vdash \mathit{Odd}$ & Made
Odd The die came up odd.	

represent it we have built a model whose verdicts are unreliable. Intuitively, this seems like a bad objection: Marilynn's information about the passage of time shouldn't affect her degrees of belief concerning the outcome of the die roll. But do our modeling framework's verdicts reflect that intuition?[9]

We respond to the objection by creating a new model D^+ (described in table 3) that represents "The odd/even announcement has been made" in its modeling language. Since our modeling language now contains sentences representing context-sensitive claims, it is possible that some sentences in the modeling language will go from a credence of 1 at t_1 to a credence less than 1 at t_2. And indeed, $\sim \mathit{Made} \in C_1^+$ but $\sim \mathit{Made} \notin C_2^+$, so $C_1^+ \not\subseteq C_2^+$. Thus we cannot derive diachronic verdicts for D^+ using (LC). However, we can obtain diachronic verdicts for D^+ by taking advantage of the relationship between D^+ and D.

D^+ is an expansion of D. To show that it is a proper expansion, it suffices to show that for every time $t_k \in T^+$ and every *atomic* sentence $y \in L^+$, there exists a sentence $x \in L$ such that $P_k^+(x \equiv y) = 1$. (Our synchronic constraints guarantee that if there is such an $x \in L$ for every *atomic* $y \in L^+$, there will be such an $x \in L$ for *every* $y \in L^+$.) Finding such equivalents for the atomic sentences of L^+ that are also atomic

9. I am not relying on this intuition because it follows from some general principle like the Relevance-Limiting Thesis. I just take it to be obvious and uncontroversial that *in this one particular case* the self-locating claim "The odd/even announcement has been made" is not relevant to Marilynn's degrees of belief about the outcome of the die roll. That makes this case a good fixed point against which to test our modeling framework.

sentences of L is easy; it is *Made* we have to worry about.[10] To show that D^+ is a proper expansion of D, we need to find an $x \in L$ such that $P_1^+(Made \equiv x) = 1$ and an $x \in L$ such that $P_2^+(Made \equiv x) = 1$.

As we noted in section 1.1, every modeling language contains at least one tautology and one contradiction. We will refer to a tautology and a contradiction in L as T and F, respectively. Then by the extrasystematic constraints on D^+ and our synchronic systematic constraints,

$$P_1^+(Made \equiv \mathsf{F}) = 1 \qquad P_2^+(Made \equiv \mathsf{T}) = 1$$

Thus D^+ is a proper expansion of D. Applying (PEP), any verdict of D has an analogue that is a verdict of D^+. In equation (1) we found that $P_2(Three) = P_1(Three \mid Odd)$ is a verdict of D, so by (PEP) we have

$$P_2^+(Three) = P_1^+(Three \mid Odd) \tag{5}$$

From there we can derive in D^+ our other intuitive verdicts for The Die. Our modeling framework reflects our intuitions about the objection: information about the passage of time is irrelevant to Marilynn's degrees of belief about the outcome of the die roll, so adding sentences representing that information to the modeling language leaves our evaluations of the requirements of ideal rationality intact.

1.6. (PEP) and Context-Sensitivity

In the previous section, we used (PEP) to fend off a challenge to verdicts about The Die we had already obtained without (PEP). But (PEP) can also yield verdicts we could not obtain otherwise. The trick is to find a proper reduction of the model in question.

Recall our model SI of Sleeping In. The modeling language of SI contains sentences representing context-sensitive claims that make $C_1 \not\subseteq C_2$ and keep us from relating P_1 to P_2 by (LC). But now consider model SI^- (described in table 4), a reduction of model SI.

We have created L^- by removing all the sentences representing context-sensitive claims from L . (The remaining sentences are meant to represent tenseless, "eternal" claims.) Since there are no context-sensitive claims represented in L^-, there are no claims represented for

10. When (now and later) I talk about an "equivalent" in L at t_1 for *Made*, I don't mean a sentence that is logically equivalent to *Made*. Instead, I mean a sentence in L representing a claim that the agent is certain at t_1 has the same truth-value as the claim represented by *Made*.

Table 4. Model SI⁻

Story: Sleeping In	1^{st} *pm* The first time Olga awakens, it is afternoon.
T^-: Contains these times:	2^{nd} *Two* The second time Olga awakens, the clock reads 2 p.m.
t_1 After Olga first awakens and sees the clock reading 6 a.m.	2^{nd} *pm* The second time Olga awakens, it is afternoon.
t_2 After Olga awakens for the second time and sees the clock reading 2 p.m.	ES: (1) $P_1^-(1^{st}Two) = 0$
	(2) $P_1^-(1^{st}pm) = 0$
	(3) $0 < P_1^-(2^{nd}Two) < 1$
L^-: Built on these atomic sentences, representing these claims:	(4) $P_2^-(2^{nd}Two) = 1$
1^{st} *Two* The first time Olga awakens, the clock reads 2 p.m.	GCS: $\langle C_2^- - C_1^- \rangle \dashv\vdash 2^{nd}Two$

which Olga goes from certainty to less-than-certainty between t_1 and t_2. Thus $C_1^- \subseteq C_2^-$, and we can use (LC) to derive:

$$P_2^-(2^{nd}pm) = P_1^-(2^{nd}pm \mid 2^{nd}Two). \tag{6}$$

We now need to demonstrate that SI⁻ is a proper reduction of SI. The following facts can be derived easily from the constraints on SI:

$$P_1(Two \equiv 1^{st}Two) = 1 \qquad P_2(Two \equiv 2^{nd}Two) = 1$$
$$P_1(pm \equiv 1^{st}pm) = 1 \qquad P_2(pm \equiv 2^{nd}pm) = 1.$$

Every atomic sentence of L not in L^- has an equivalent in L^- at each time in T. As we noted in section 1.5, this is sufficient to establish that SI⁻ is a proper reduction of SI. So we can apply (PEP) to derive the following verdict of SI from equation (6):

$$P_2(2^{nd}pm) = P_1(2^{nd}pm \mid 2^{nd}Two). \tag{7}$$

One important consequence of our synchronic constraints is a principle I'll call **substitution**: if $P_i(x \equiv y) = 1$, y can be replaced in a verdict with x anywhere it appears in a P_i expression. By substitution and the fact that $P_2(pm \equiv 2^{nd}pm) = 1$, equation (7) becomes

$$P_2(pm) = P_1(2^{nd}pm \mid 2^{nd}Two). \tag{8}$$

Equation (8) answers the question originally posed in Sleeping In: it relates Olga's degree of belief at t_2 that it is then afternoon to

her degrees of belief at t_1. Intuitively, it also represents a requirement of ideal rationality. Imagine that as she drifts off to sleep at t_1, Olga asks herself "How confident am I that my next awakening will occur during the afternoon, conditional on the supposition that on that awakening my clock will read 2 p.m.?" Whatever her answer is, when Olga actually awakens the second time, sees that her clock reads 2 p.m. and asks herself how confident she is that it is *now* afternoon, her answer should be the same. Having rejected Conditionalization as a systematic constraint and replaced it with (LC) and (PEP), we can now derive a verdict capturing this intuition.

How exactly did we derive that verdict? The presence of sentences representing context-sensitive claims in L allows C_1 not to be a subset of C_2. Because this subset relation fails, we cannot apply (LC) to SI to generate diachronic verdicts. But we can take advantage of another feature of L's structure: for each context-sensitive claim represented in L and each time in the time set, there is a context-insensitive claim represented in L that Olga is certain at that time has the same truth-value as the context-sensitive claim.

We can build a modeling language with this structure because at each time in the time set there is a context-insensitive expression that Olga is certain uniquely picks out the denotation of the context-sensitive expression "now" at that time. The context-sensitive claims represented in L are context-sensitive because of the shifting denotation of "now." But at t_1 Olga is certain that "now" picks out the same time as "the first time Olga awakens," and at t_2 Olga is certain that "now" picks out the same time as "the second time Olga awakens." At each time, Olga can substitute the relevant context-insensitive expression for "now" into each context-sensitive claim, yielding a context-insensitive claim that she is certain has the same truth-value.

The key feature of L is that, while it contains sentences representing context-sensitive claims, it also contains sentences representing the context-insensitive equivalents of those claims that result from such substitutions. For example, substituting "the first time Olga awakens" for "now" in "The clock now reads 2 p.m." yields "The first time Olga awakens, the clock reads 2 p.m." Both the former claim and the latter claim are represented by sentences of L, and at t_1 Olga is certain that the two have the same truth-value. (I should say that ideal rationality *requires* Olga to be certain that the two have the same truth-value, but I will omit that qualification from now on.)

For each time in the time set of SI and each context-sensitive claim represented in its modeling language, there is a context-insensitive claim represented in that language that Olga is certain has the same truth-value at that time. We can therefore construct a proper reduction SI^- of SI whose modeling language represents only the context-insensitive claims represented in L. Because L^- represents only context-insensitive claims, there are no claims that go from certainty to less-than-certainty in SI^-. So while (LC) could not be usefully applied to SI, it can be applied to SI^- to relate Olga's t_1 and t_2 degrees of belief. Since SI is a proper expansion of SI^-, (PEP) then guarantees that the analogues of verdicts of SI^- will also be verdicts of SI. Finally, since Olga is certain at t_2 that "It is now afternoon" has the same truth-value as "The second time Olga awakens, it is afternoon," a verdict originally derived in SI^- relating the latter to Olga's first-awakening degrees of belief can be used in SI to relate the former to those degrees of belief. This yields our result, equation (8).

1.7. The Rationale for (PEP)

We have just seen that the addition of (PEP) to our modeling framework allows its models to yield verdicts they would not yield otherwise. But why should we believe that these verdicts will always represent requirements of ideal rationality?

The best way to evaluate (PEP) is to evaluate the entire modeling framework of which it forms a part. To do so, we test the framework on stories in which the requirements of ideal rationality are intuitively obvious and agreed-upon by all. For each such story on which I have tested our framework—including The Die, Sleeping In, and others—I have obtained verdicts that clearly represent requirements of ideal rationality. This strikes me as the best support for our framework as a whole and for (PEP)'s role within it.

Still, there are further things that can be said in favor of (PEP). For example, there are particular instances of (PEP) that can be proven to follow from our other systematic constraints. Corollary (A.7) in appendix A shows that if the model M^+ referred to in (PEP) contains no sentences that go from a credence of 1 to a credence less than 1, (PEP) is a theorem provable from our other systematic constraints. So when working with an M^+ whose modeling language represents no context-sensitive claims, we can be as confident of verdicts derived using (PEP) as we are of verdicts derived solely from our other constraints.

Table 5. Atomic Sentences in Models B and B^+

L:		L^+:	
GotJoker	The Joker is apprehended.	*GotJoker*	The Joker is apprehended.
BatFinds	Batman finds the Joker's hideout.	*BatFinds*	Batman finds the Joker's hideout.
		BruceBat	Bruce Wayne is Batman.
		BruceFinds	Bruce Wayne finds the Joker's hideout.

Corollary (A.4) shows that if M^+ is an expansion of M, any verdict of M that can be derived solely from extrasystematic constraints and our *synchronic* systematic constraints has an analogue in M^+ that can be derived from extrasystematic constraints and synchronic constraints. Thus even when context-sensitive claims are represented in the language of M^+, verdicts of (PEP) whose derivations involve only one time can be shown to follow from our other constraints. Given these results, we can read (PEP) as extending to diachronic, context-sensitive cases a principle that follows from our other systematic constraints for both context-insensitive cases and synchronic cases.

One can get a further sense of what (PEP) does by considering two simple examples. First, imagine a model B and its expansion B^+ whose modeling languages have the atomic sentences listed in table 5. Suppose that between t_1 and t_2 (the only times in T) our agent becomes certain of the claim represented by *BatFinds*. We can use B to model this information's effects on the agent's credence in *GotJoker*.

Now suppose that throughout the story the agent is certain of the claim represented by *BruceBat*. There are no context-sensitive claims represented in L^+, and therefore no claims going from certainty to less-than-certainty in B^+. Thus by corollary (A.7), it follows from our first five systematic constraints that analogues of B's verdicts will be verdicts of B^+. In this situation, adding sentences representing claims about Bruce Wayne to the modeling language does not alter verdicts about how the agent's credence in *GotJoker* will respond to what she learns. And since B^+ is a proper expansion of B, this is just what (PEP) would predict.

Why does moving to B^+ leave B's verdicts intact? In B^+, the agent becomes certain between t_1 and t_2 not only of the claim represented by *BatFinds*, but also of the claim represented by *BruceFinds*. But since the agent is certain throughout the story that Bruce Wayne is Batman, *BruceFinds* gives her no further information relevant to *GotJoker* beyond

Table 6. Atomic Sentences in Models W and W^+

L:		L^+:	
Gases	Greenhouse gases raise the Earth's temperature.	*Gases*	Greenhouse gases raise the Earth's temperature.
4^{th} *Hot*	July 4th, 2006 is unseasonably hot.	4^{th} *Hot*	July 4th, 2006 is unseasonably hot.
5^{th} *Hot*	July 5th, 2006 is unseasonably hot.	5^{th} *Hot*	July 5th, 2006 is unseasonably hot.
		4^{th}	Today is July 4th, 2006.
		5^{th}	Today is July 5th, 2006.
		Hot	Today is unseasonably hot.

what was contained in *BatFinds*. As for *BruceBat*, while the fact that Bruce Wayne is Batman is certainly significant from a broader perspective, *from the point of view of the modeling framework* the claim represented by *BruceBat* is merely a piece of linguistic information—it gives the agent a *synonym* for "Batman." With *BruceBat* in place, the agent can describe what she believes using L^+ "Bruce Wayne" claims that she is certain have the same truth-values as L's "Batman" claims. But this new linguistic ability does not give her any additional information relevant to *GotJoker*.

Now consider another example. Suppose model W^+ is an expansion of model W with time set $\{t_1, t_2\}$, and the atomic sentences of their modeling languages are as described in table 6. Suppose further that at t_1 the agent is certain of 4^{th} & $\sim 5^{th}$ and at t_2 the agent is certain of $\sim 4^{th}$ & 5^{th}. Then W will be a proper reduction of W^+, and (PEP) ensures that whatever W tells us about the agent's credence in *Gases* will be borne out by W^+.

Why should W's verdicts be maintained in W^+? L^+ represents context-sensitive claims (such as 4^{th}) that go from certainty to less-than-certainty between t_1 and t_2, so appendix A's results will not underwrite (PEP)'s diachronic verdicts here. But the idea from the Batman example still applies: the extra claims in L^+ have simply added a synonym, "today," that can be used to reexpress claims in L. Because of its context-sensitivity, the synonym "today" behaves a bit curiously; it is synonymous with one context-insensitive expression ("July 4th, 2006") at t_1 and with another ("July 5th, 2006") at t_2. From a broader philosophy of language point of view it may be misleading to think of "today" as a synonym here, but *from the point of view of the modeling framework* the extra linguistic information in L^+ still just gives the agent new ways of saying the same old

things; no additional information has been represented that is relevant to *Gases*. W^+ should maintain W's verdicts, and this is exactly what (PEP) achieves.

One might object that if Bayesianism has taught us anything, it's that almost anything can be relevant to almost anything else. Despite the fact that 4^{th} and 5^{th} represent mere "linguistic information," there must be some way to arrange a story so that the move from W to W^+ alters verdicts about *Gases*. But this objection, if sound, should apply to the Batman example as well. And there we can prove from our first five systematic constraints that there is no story in which adding linguistic information about "Bruce Wayne" to the modeling language alters our original verdicts.

Ultimately, I do not want to rest too much weight on these arguments. As I said above, the best support for (PEP) is the successful application of our modeling framework to stories in which the requirements of ideal rationality are uncontroversial. The main point of the Batman and greenhouse examples is to give the reader an intuitive sense of what (PEP) does.

One final note: with (PEP) in place, we can respond to the objection raised in section 1.4 that the restriction on conditionalizing in (LC) is an overreaction. Suppose that instead of (LC) we choose for our diachronic constraint a less-limited conditionalization principle—call it (LLC)—that declines to conditionalize only when some sentence in the modeling language goes from one extremal unconditional credence to the other between two times. Now consider the reduction SI* of SI whose modeling language L^* contains the atomic sentences *pm*, 1^{st}*Two*, 1^{st}*pm*, 2^{nd}*Two*, and 2^{nd}*pm*. The reader can work out that in L^* there is no sentence whose credence goes from one extreme to the other between t_1 and t_2. Thus (LLC) would apply to SI* to yield $P_2^*(pm) = 0$. SI* is a proper reduction of SI, so by (PEP) we would have $P_2(pm) = 0$, a verdict we do not want. But using (LC) as our diachronic constraint prevents this result because there are sentences in L^* that go from certainty to less-than-certainty between t_1 and t_2. So (LC) yields no diachronic verdicts for SI*.

1.8. The Remaining Problem

Even with unimpeachable systematic and extrasystematic constraints, a model may yield verdicts that fail to represent requirements of ideal rationality if its modeling language is impoverished—that is, if the

modeling language fails to represent some relevant claims. We recognize this threat with a modeling rule: if we have a model and its expansion, and the analogues of the model's verdicts are not verdicts of the expansion, we should not trust the original model's verdicts to represent requirements of ideal rationality.

This can happen in one of two ways. First, the original model may yield a verdict whose analogue *contradicts* a verdict of the expansion. For example, if we take model D of The Die and construct a reduction D^- whose only atomic sentence is *Three*, the gained certainty set in that model is empty, and (LC) yields the verdict $P_2^-(\mathit{Three}) = P_1^-(\mathit{Three})$. The analogue of this verdict in D contradicts D's verdict that $P_2(\mathit{Three}) > P_1(\mathit{Three})$. This suggests that the verdicts of D^- may not represent requirements of ideal rationality, a possibility that is borne out by our intuitions about The Die.[11]

Second, the original model may yield a verdict whose analogue is not contradicted by the expansion, but is also not a verdict of the expansion. Again, the original model's verdicts are not to be trusted. For example, suppose that in some model M, (LC) applies to yield diachronic verdicts. But suppose an improper expansion of M, M^+, represents context-sensitive claims in its modeling language that are not represented in the modeling language of M. If any of these extra claims goes from certainty to less-than-certainty during the story, (LC) will fail to yield any diachronic verdicts for M^+. Here the expansion does not yield verdicts *contradicting* M's diachronic verdicts, but because it also fails to replicate M's diachronic verdicts we should not rely on those verdicts to represent requirements of ideal rationality. Instead, we should try to obtain diachronic verdicts for M^+ via some other route. If the modeling language of M^+ is structured like the modeling language of SI—that is, if it contains context-insensitive truth-value equivalents for each context-sensitive sentence at each time in the time set—we can construct a reduction of M^+ different from M whose language represents only the context-insensitive claims represented in M^+. This model will yield diachronic verdicts by (LC), and since it is a proper reduction of M^+, those verdicts can be exported back to M^+ by (PEP). If the modeling language of M^+ does not have the necessary structure, we may have

11. Notice that D^- does not provide a counterexample to (PEP) because D^- is not a *proper* reduction of D.

to construct a further expansion M^{++} of M^{+} whose modeling language does.[12]

Ultimately we should trust the verdicts of the model whose language is a superset of the languages of all the models we have tried for a story. This modeling rule addresses the concern expressed in section 1.4 that moving from Conditionalization to (LC) is not enough to solve Conditionalization's problems—that (LC) will still yield verdicts that do not represent requirements of ideal rationality. We might have a story in which ideal rationality requires an agent to change her degree of belief in some claim between two times, but when we examine our model we find no sentences going either from less-than-certainty to certainty or vice versa between those times. The agent's certainty set will not have changed, so (LC) will yield a verdict requiring her degrees of belief to remain identical from the earlier to the later time.[13] In this case, we should consider whether there is a claim not represented in our model's language that goes from less-than-certainty to certainty or vice versa between the two times. If there is, we can represent that claim in the modeling language of an expansion of our original model, an expansion which will not yield verdicts requiring the agent's degrees of belief to remain fixed. A fault that seemed initially to lie in (LC) turns out to lie in our choice of modeling language.

There may be stories in which ideal rationality requires an agent to alter her degrees of belief between two times despite the fact that *no* claims become certain or lose certainty between those two times. If there are such stories, they lie outside the domain that can be accurately modeled using our modeling framework. As I stated when I first defined a "story" in section 1.1, our framework is designed to model the effects of evolving certainty sets on an agent's nonextreme degrees of belief. Stories in which rationality requires degrees of belief to change without *any* changes in the agent's certainties are better modeled using an alternative approach, perhaps one based on Jeffrey Conditionalization.

Yet even within its intended domain, our modeling framework still faces a problem. The strategy we have presented so far for modeling stories involving context-sensitive claims that go from certainty to less-than-certainty requires us to construct a modeling language

12. I am grateful to Peter Vranas for pressing me to address the case considered in this paragraph.

13. I am grateful to an anonymous referee for the *Philosophical Review* for raising this concern.

structured like the language of SI. That structure allows us to move down to a context-insensitive proper reduction whose diachronic verdicts can be brought back up by applying (PEP). In order to construct a modeling language with that structure, we need it to be the case that for each time in the time set and each context-sensitive claim represented in the language, there is a context-insensitive expression that the agent is certain at that time uniquely picks out the denotation of the context-sensitive expression in the context-sensitive claim.[14] But what if, due to the vagaries of the story, our agent lacks a uniquely denoting context-insensitive expression for a context-sensitive expression at some time? We need a strategy for applying our modeling framework to stories in which this occurs.

This problem is particularly relevant to our attack on the Relevance-Limiting Thesis. Any counterexample to the thesis must be a story in which an agent becomes certain of some self-locating claims between two times without becoming certain of any non-self-locating claims. Our counterexample, the Sleeping Beauty Problem, achieves this by ensuring that the agent lacks a uniquely denoting context-insensitive expression for a context-sensitive expression at a particular time. So to derive diachronic verdicts for the Sleeping Beauty Problem, we need a strategy for analyzing such stories using our modeling framework. Developing such a strategy will be the focus of this essay's second half.

2. The Sleeping Beauty Problem

2.1. The Problem

> *The Sleeping Beauty Problem:* A student named Beauty volunteers for an on-campus experiment in epistemology. She arrives at the lab on Sunday, and the details of the experiment are explained to her in full. She will be put to sleep Sunday night; the experimenters will then flip a fair coin. If the coin comes up heads, they will awaken her Monday morning, chat with

14. The context-sensitive claim may contain more than one context-sensitive expression, in which case we need the agent to have a uniquely denoting context-insensitive expression for *each* context-sensitive expression at each time. (Clearly we are concerned only with context-sensitive expressions occurring in nonintensional contexts.)

> her for a bit, then put her back to sleep. If the coin comes up tails, they will engage in the same Monday process then *erase all her memories of her Monday awakening*, awaken her Tuesday morning, chat with her for a bit, then put her back to sleep.
>
> Beauty is told and believes with certainty all the information in the preceding paragraph, then she is put to sleep. Some time later she finds herself awake, uncertain whether it is Monday or Tuesday. What does ideal rationality require at that moment of Beauty's degree of belief that the coin came up heads?

Adam Elga (2000) argues that Beauty's degree of belief in heads should be one-third, while Lewis (2001) argues it should be one-half.

In analyzing this story, I will assume that Beauty's Monday and Tuesday awakenings are subjectively indistinguishable. (Section 2.5 considers a version of the story in which they are not.) Further, I will assume that Beauty remains in the same room throughout the experiment and studies it intently enough on Sunday night that she gains no non-self-locating information about her surroundings when she awakens Monday morning. With these assumptions in place, Beauty learns no non-self-locating claims between Sunday night and Monday morning. If ideal rationality nevertheless requires her to change her degree of belief between Sunday night and Monday morning in the non-self-locating claim that the coin comes up heads, the Sleeping Beauty Problem provides a counterexample to the Relevance-Limiting Thesis.

Elga analyzes the Sleeping Beauty Problem by adding a feature to the story. He imagines that, as part of the experimental protocol, on each day that Beauty is awakened, the researchers chat with her for a bit and then reveal to her what day it is before putting her back to sleep. (If Beauty's memories of her Monday awakening are erased, this revelation is among the information lost.) The additional part of the protocol is explained to Beauty on Sunday, so she is certain that each time she awakens she will eventually be told what day it is.

We can model Beauty's Monday reaction to learning what day it is using model SB1, described in table 7. Note that in this model *Heads* represents a tenseless claim. Extrasystematic constraint (3) comes from

Table 7. Model SB1

Story: Sleeping Beauty	*L*: Built on these atomic sentences, representing these claims:
T: Contains these times:	*Monday* Today is Monday.
t_1 Monday morning, after Beauty awakens but before she is told what day it is.	*Heads* The coin comes up heads.
t_2 Monday night, after Beauty has been told it is Monday but before she is put back to sleep.	ES: (1) $0 < P_1(Monday) < 1$ (2) $P_2(Monday) = 1$ (3) $P_1(Heads \supset Monday) = 1$
	GCS: $\langle C_2 - C_1 \rangle \dashv\vdash Monday$

the structure of the experiment: if the coin comes up heads, Beauty is awakened only on Monday.

L contains sentences (such as *Monday*) representing context-sensitive claims. But between t_1 and t_2, "Today is Monday" does not change its truth-value. Moreover, at both t_1 and t_2 Beauty is certain that "Today is Monday" has the same truth-value during the morning of the current day as it has during the evening (even though at t_1 Beauty is not certain what that truth-value is). Thus *Monday* behaves in SB1 like a sentence representing a context-insensitive claim. This is reflected in the fact that no sentences in *L* go from certainty to less-than-certainty between t_1 and t_2. So we can apply (LC) to derive:

$$P_2(Heads) = P_1(Heads \mid Monday). \tag{9}$$

Applying our synchronic systematic constraints and extrasystematic constraint (3) yields

$$P_2(Heads) = \frac{P_1(Heads \,\&\, Monday)}{P_1(Monday)} = \frac{P_1(Heads)}{P_1(Monday)}. \tag{10}$$

By extrasystematic constraint (1), the denominator is less than 1, so

$$P_2(Heads) > P_1(Heads). \tag{11}$$

Equation (11) makes sense intuitively. $P_1(Heads)$ is a weighted average of the values $P_1(Heads \mid Monday)$ and $P_1(Heads \mid \sim Monday)$. Since the latter value is 0 (if Beauty is awake on Tuesday, the coin came up tails), $P_1(Heads)$ must be less than $P_1(Heads \mid Monday)$. By equation (9), $P_2(Heads) = P_1(Heads \mid Monday)$, so $P_2(Heads) > P_1(Heads)$.

Though Lewis and Elga analyze the Sleeping Beauty Problem using a different modeling framework than ours, both of them agree with the results obtained so far. From this point their arguments diverge.

Elga applies the Principal Principle to determine directly the value of $P_2(Heads)$. For our purposes, we can take the Principal Principle to say that if an agent with no inadmissible evidence is certain a particular outcome of a chance process has a particular objective chance, ideal rationality requires her to set her degree of belief in that outcome equal to that chance. Inadmissible evidence is evidence that indicates how the chance process came out; it influences an agent's degree of belief that a particular outcome has occurred without influencing her beliefs about that outcome's objective chance.[15]

On Sunday night, Beauty is certain that the coin is fair and she has no evidence about the outcome of the flip. Every party to the Sleeping Beauty debate agrees that at that point the Principal Principle requires her to assign degree of belief one-half to heads. Elga notes that since the experimenters are going to awaken Beauty on Monday whether the coin comes up heads or tails, it makes no difference to the experimental protocol if the coin is flipped after Beauty goes to sleep Sunday night or after she goes to sleep on Monday. If the coin is flipped Monday night, the Principal Principle seems to require Beauty to assign $P_2(Heads) = 1/2$: at t_2 Beauty is still certain that the coin is fair and hasn't been flipped yet, so how could she have inadmissible evidence about its outcome? Given his $P_2(Heads)$ assignment for the Monday-flip case, and the fact that flipping the coin on Monday makes no difference to the experimental protocol, Elga argues that Beauty is also required to assign $P_2(Heads) = 1/2$ in the Sunday-flip case.

Elga (2000, 144) then applies a "highly restricted principle of indifference" to assign $P_1(Monday \mid \sim Heads) = 1/2$. His thought is that on Monday morning when Beauty is uncertain what day it is, she should be equally confident that it is Monday or Tuesday on the supposition that the coin came up tails. With his $P_2(Heads)$ value, his $P_1(Monday \mid \sim Heads)$ value, and $P_1(Monday \mid Heads) = 1$ from the story, Elga applies Bayes's Theorem to equation (9) and calculates

15. Technically evidence is admissible or inadmissible for an agent only relative to a particular time and outcome. To simplify locutions I will typically leave those qualifiers implicit in our discussion. (I am grateful to an anonymous referee for the *Philosophical Review* for raising this concern.)

$P_1(Heads) = 1/3$. Elga concludes that when Beauty first awakens she should assign a degree of belief of one-third to heads.

Lewis, on the other hand, argues directly to a $P_1(Heads)$ value from three premises. With Lewis's notation changed to match ours (and t_0 representing Sunday night before Beauty goes to sleep), they are:

- $P_0(Heads) = 1/2$
- "Beauty gains no new uncentred evidence, relevant to Heads versus Tails, between the time when she has credence function P_0 and the time when she has credence function P_1. The only evidence she gains is the centred evidence that she is presently undergoing either the Monday awakening or the Tuesday awakening: that is, ['Today is Monday or Tuesday']." (Lewis 2001, 173)
- "Only new relevant evidence, centred or uncentred, produces a change in credence; and the evidence ['Today is Monday or Tuesday'] is not relevant to Heads versus Tails." (ibid. 174)

From these premises it follows that $P_1(Heads) = 1/2$: Beauty should assign the same degree of belief to heads on Monday morning that she did on Sunday night.

Lewis never says *why* he thinks "Today is Monday or Tuesday" is not relevant to heads. The idea may be that since Beauty was already certain on Sunday night she was going to awaken in the experimenters' room during the week, when she finally awakens in that room, her new self-locating evidence that "Today is Monday or Tuesday" is relevant only to self-locating degrees of belief. This is certainly the argument made most often in conversation by defenders of Lewis's position.

But the Lewis defender cannot support this argument by a principled appeal to the Relevance-Limiting Thesis because Lewis's position is inconsistent with that thesis.[16] Lewis accepts equation (11); he agrees with Elga that Beauty's degree of belief in heads should increase between Monday morning and Monday night. Yet the only claims Beauty learns between those two times are self-locating, so by granting that Beauty is rationally required to alter her non-self-locating degree of belief in heads between t_1 and t_2, Lewis contravenes the Relevance-Limiting Thesis. To maintain that thesis consistently, one would have to

16. I am grateful to Darren Bradley for pointing this out to me. A similar point is made at Bostrom 2007, 66.

Table 8. Model SB0

Story: Sleeping Beauty	*L*: Built on these atomic sentences, representing these claims:
T: Contains these times:	*Monday* Today is Monday.
t_0 Sunday night, after Beauty has heard the experiment described but before she is put to sleep.	*Heads* The coin comes up heads.
t_2 Monday night, after Beauty has been told it is Monday but before she is put back to sleep.	ES: (1) $P_0(Monday) = 0$ (2) $P_2(Monday) = 1$ (3) $0 < P_0(Heads) < 1$ (4) $0 < P_2(Heads) < 1$
	GCS: $\langle C_2 - C_0 \rangle \dashv\vdash Monday$

argue that Beauty's degree of belief in heads should remain constant at one-half from t_0 to t_1 to t_2. In sections 2.3 and 2.7 we will consider positions which argue exactly that.

Because he holds that $P_2(Heads) > 1/2$, Lewis also has to argue (*contra* Elga) that on Monday night Beauty possesses inadmissible evidence concerning the coin flip and so is permitted by the Principal Principle to deviate her degree of belief in heads from one-half. If we had a precise, general procedure for determining when evidence is admissible for particular claims, we could use the Principal Principle to adjudicate this disagreement over Beauty's ideally rational Monday night degree of belief in heads. But another approach is available: by applying the modeling framework developed in the first half of this essay, we can refute Lewis's position without appealing to the Principal Principle at all.

2.2. The Solution

To complete our analysis of the Sleeping Beauty Problem, we need a model representing Sunday night and Monday night. Such a model, SB0, is described in table 8.

L contains sentences (such as *Monday*) representing context-sensitive claims that go from one extremal degree of belief to the other between t_0 and t_2. So (LC) cannot yield diachronic verdicts for this model. However, the context-sensitivity of these claims derives from the context-sensitivity of "today," and Beauty has a uniquely denoting context-insensitive expression for "today" at both t_0 and t_2. At t_0 she can replace "today" with "Sunday" in the claim "Today is Monday," yielding a contradiction; at t_2 she can replace "today" with "Monday," yielding a

Table 9. Model SB0$^-$

Story: Sleeping Beauty	L^-: Built on this atomic sentence, representing this claim:
T: Contains these times:	*Heads* The coin comes up heads.
t_0 Sunday night, after Beauty has heard the experiment described but before she is put to sleep.	ES: (1) $0 < P_0^-(\mathit{Heads}) < 1$ (2) $0 < P_2^-(\mathit{Heads}) < 1$
t_2 Monday night, after Beauty has been told it is Monday but before she is put back to sleep.	CGS: $\langle C_2^- - C_0^- \rangle \dashv\vdash \mathsf{T}$

tautology. So instead of working with SB0, we can work with a reduction whose only atomic sentence is *Heads*. This reduction, model SB0$^-$, is described in table 9.

There are no context-sensitive claims represented in L^-, and therefore no sentences going from certainty to less-than-certainty. So we can derive verdicts using (LC), and since $C_2^- = C_0^-$, we have:

$$P_2^-(\mathit{Heads}) = P_0^-(\mathit{Heads}). \tag{12}$$

Moreover, since $P_0(\mathit{Monday} \equiv \mathsf{F}) = 1$ and $P_2(\mathit{Monday} \equiv \mathsf{T}) = 1$ (with T and F representing a tautology and a contradiction in L^-), SB0$^-$ is a proper reduction of SB0. Applying (PEP),

$$P_2(\mathit{Heads}) = P_0(\mathit{Heads}). \tag{13}$$

In the previous section, our analysis of model SB1 showed that ideal rationality requires Beauty's Monday morning degree of belief in heads to be less than her Monday night degree of belief in heads. Our analysis of model SB0 now shows that ideal rationality requires Beauty's Monday night degree of belief in heads to equal her Sunday night degree of belief in heads. Thus ideal rationality requires Beauty's Monday morning degree of belief in heads to be less than her Sunday night degree of belief in heads.

If we wanted, we could appeal to the Principal Principle and place an extrasystematic constraint on SB0 that $P_0(\mathit{Heads}) = 1/2$, allowing us eventually to conclude that $P_1(\mathit{Heads}) < 1/2$. But notice that all the conclusions in the previous paragraph were derived strictly

through analyses of SB1 and SB0 using our framework's systematic constraints; neither the Principal Principle nor any indifference principle was required. Simply by applying a framework that properly models the effects of context-sensitive claims on ideally rational degrees of belief, we can show that ideal rationality requires Beauty's Monday morning degree of belief in heads to be less than her Sunday night degree of belief in heads. And that is sufficient to refute Lewis's position.

2.3. Objections to This Solution

Nick Bostrom (2007) defends a solution to the Sleeping Beauty Problem on which $P_1(\mathit{Heads}) = P_2(\mathit{Heads}) = 1/2$. He makes this plausible by rejecting the conditionalizing step that yielded $P_2(\mathit{Heads}) = P_1(\mathit{Heads} \mid \mathit{Monday})$ (equation (9)), noting that model SB1 fails to represent a claim Beauty learns between Monday morning and Monday night. Between t_1 and t_2 Beauty becomes certain not only of the claim "Today is Monday," but also of the claim "I have been told today that today is Monday." Bostrom argues that Beauty's Monday night degree of belief in heads is required to equal her Monday morning degree of belief in heads conditional on *both* these claims, and there is no reason to think that this conditional degree of belief should be greater than Beauty's unconditional Monday morning degree of belief in heads. So we cannot rely on equation (11)'s conclusion that $P_2(\mathit{Heads}) > P_1(\mathit{Heads})$.

This objection is easily evaluated using (PEP). We construct a model SB1$^+$ (whose full description I leave to the reader) that adds to the modeling language of SB1 the sentence *Told*, representing "I have been told today that today is Monday." Since Beauty is certain at t_1 that this claim is false, we have $P_1^+(\mathit{Told} \equiv \mathsf{F}) = 1$ for any contradiction $\mathsf{F} \in L$; since Beauty is certain at t_2 that the claim is true we have $P_2^+(\mathit{Told} \equiv \mathsf{T})$ for any tautology $\mathsf{T} \in L$. So SB1$^+$ will be a proper expansion of SB1. By (PEP), analogues of the verdicts of SB1 (in particular equations (9), (10), and (11)) will be verdicts of SB1$^+$. SB1$^+$ will therefore yield the verdict that Beauty's Monday night degree of belief in heads is greater than her Monday morning degree of belief in heads. Taking "I have been told today that today is Monday" into account makes no difference to our verdicts about Beauty's degrees of belief in heads.[17]

17. Since Beauty is certain that each time she awakens she will eventually be told what day it is, incorporating "I have been told today that today is Monday" into our model is just incorporating information about the passage of time; all it adds to "Today is Monday" is that the time when the announcement is made has passed. As a result,

Taking a different approach, one might object that our analysis ignores the role of memory loss in the Sleeping Beauty Problem. Frank Arntzenius (2003) has described some examples in which the presence of forgetting causes Conditionalization to yield verdicts that do not represent requirements of ideal rationality. Even worse, Conditionalization can yield incorrect verdicts when an agent merely *suspects* she has forgotten information since an earlier time, even if no memory loss has actually occurred. In the Sleeping Beauty Problem, the experimenters do not tamper with Beauty's memory between Sunday night and Monday morning. But on Monday morning Beauty suspects it might be Tuesday and so cannot be certain that none of her memories has been erased. Thus we might worry that (LC) is an unreliable tool for analyzing Beauty's Monday morning degrees of belief.

Conditionalization fails when forgetting or the threat of forgetting causes an agent to go from certainty in a claim to less-than-certainty in that claim, as in Arntzenius's examples. In the Sleeping Beauty Problem, Beauty goes from certainty to less-than-certainty in various claims between Sunday night and Monday morning, but that is due exclusively to those claims' context-sensitivity. Were memory erasure not a part of the experimental protocol, Beauty would be certain on Monday morning that it is her first awakening, and so become certain of various claims (such as "Today is Monday") that she is not certain of in the actual problem. The threat of memory loss prevents Beauty from *forming* various certainties on Monday morning that she might have formed otherwise, but it does not cause her to *lose* any certainties she had on Sunday night. So we can rely on (LC) in analyzing Beauty's Monday morning degrees of belief.[18]

In fact, the memory erasure in the Sleeping Beauty Problem is significant only because it leaves Beauty without a uniquely denoting context-insensitive expression for a significant context-sensitive expression. If it weren't for the threat of memory loss, Beauty's awakenings would be subjectively distinguishable because on Monday morning she could describe "today" as "the first day on which I awaken during the experiment." We can create stories that are like the Sleeping Beauty Problem but involve no memory loss by substituting another device that

---

our analysis of $SB1^+$ precisely mirrors our analysis of model D^+ in section 1.5, where we demonstrated the irrelevance of similar information about the passage of time.

18. In Titelbaum 2008 I show that Arntzenius's examples can be properly modeled using (LC), as can similar examples in Talbott 1991.

leaves Beauty without a uniquely denoting context-insensitive expression for a significant context-sensitive expression. For example, instead of awakening Beauty twice if the coin lands tails, the experimenters could awaken her only once but make a perfect copy of her and awaken it at the same time in an indistinguishable room.[19] Instead of being uncertain what day it is, and so lacking a uniquely denoting context-insensitive expression for the context-sensitive expression "today," Beauty would be uncertain whether she is Beauty or the doppelgänger and so would lack a uniquely denoting context-insensitive expression for the context-sensitive expression "I." The resulting doppelgänger story is structurally identical to the original Sleeping Beauty Problem.

2.4. Modeling Strategy

Section 1.6 described a strategy for deriving diachronic verdicts for models whose languages include sentences representing context-sensitive claims. That strategy requires the modeling language to contain context-insensitive equivalents for each context-sensitive claim represented. This, in turn, requires the agent to have a uniquely denoting context-insensitive expression at each time for each context-sensitive expression in a claim represented in the modeling language.

The original Sleeping Beauty Problem concerns two times: Sunday night and Monday morning. Between those two times, Beauty becomes certain of the claim "Today is Monday or Tuesday" and loses certainty in the claim "Today is Sunday." Though Beauty is uncertain on Monday morning what day it is, she is certain those claims have changed their truth-values since Sunday night. And because she is uncertain what day it is, she lacks a context-insensitive expression that she can be certain uniquely picks out the denotation of "today."

How, then, did we relate Beauty's Sunday night degrees of belief to her Monday morning degrees of belief using our modeling framework? By adding a feature to the story. In sections 2.1 and 2.2 we followed Elga and added to the story a time on Monday night when Beauty is certain what day it is. We then related Beauty's Sunday night degrees of belief to her Monday morning degrees of belief indirectly, in two steps. First, we related Monday morning to Monday night using model SB1. Because Beauty is certain claims containing "today" do not change

19. Doppelgänger stories like this appear in Arntzenius 2003, Bostrom 2007, Elga 2004, and Meacham 2008.

their truth-values between those two times, we could use (LC) to derive diachronic verdicts for SB1. Second, we related Sunday night to Monday night using model SB0. Claims containing "today" do change their truth-values between Sunday night and Monday night, but at each of those two times Beauty has a uniquely denoting context-insensitive expression for "today." So we could construct a proper reduction of SB0 (model $SB0^{-}$) whose language represented only context-insensitive claims, then use (LC) to derive diachronic verdicts for it. Combining our results from these models, we related Beauty's Sunday night degrees of belief to her degrees of belief on Monday morning.

When we add a feature to a story for modeling purposes, we must be careful that the added feature does not alter the relations we hoped to model in the first place. Elga assumes that adding a time *after* Monday morning at which Beauty is told what day it is does not alter her required degree of belief in heads *on* Monday morning. Our solution relies on this assumption as well.

The next section presents an alternative solution to the Sleeping Beauty Problem that relates Beauty's Sunday night degrees of belief to her Monday morning degrees of belief directly. We have been assuming that Beauty's Monday morning and Tuesday morning awakenings are subjectively indistinguishable. But we can add a feature to the story that allows Beauty to distinguish the two awakenings. The trick is to keep this addition independent of the degrees of belief we are after; we must be confident that our addition does not alter the requirements of ideal rationality on Beauty's Monday morning degree of belief in heads.

2.5. Technicolor Beauty

> *Technicolor Beauty:* Everything is exactly as in the original Sleeping Beauty Problem, with one addition: Beauty has a friend on the experimental team, and before she falls asleep Sunday night he agrees to do her a favor. While the other experimenters flip their fateful coin, Beauty's friend will go into another room and roll a fair die. (The outcome of the die roll is independent of the outcome of the coin flip.) If the die roll comes out odd, Beauty's friend will place a piece of red paper where Beauty is sure to see it when she awakens Monday

> morning, then replace it Tuesday morning with a blue paper she is sure to see if she awakens on Tuesday. If the die roll comes out even, the process will be the same, but Beauty will see the blue paper on Monday and the red paper if she awakens on Tuesday.
>
> Certain that her friend will carry out these instructions, Beauty falls asleep Sunday night. Some time later she finds herself awake, uncertain whether it is Monday or Tuesday, but staring at a colored piece of paper. What does ideal rationality require at that moment of Beauty's degree of belief that the coin came up heads?

To simplify discussion, we will focus on the case in which Beauty awakens to a red piece of paper on Monday; this choice is made without loss of generality and our analysis would proceed identically for the blue-Monday case. We will analyze Technicolor Beauty using model TB, described in table 10. Note that in this model *Heads*, *MonRed*, and *UpRed* represent tenseless claims. (The extrasystematic constraints on TB are explained in appendix B.)

Model TB aims to directly relate Beauty's Sunday night and Monday morning degrees of belief. But its modeling language contains sentences (such as $\sim$*Monday*) that go from certainty to less-than-certainty between t_0 and t_1. Thus (LC) cannot yield diachronic verdicts for TB. However, the addition of the colored papers has given Beauty a uniquely denoting context-insensitive expression for "today." On Monday morning, Beauty is certain that "the red paper day" uniquely picks out the denotation of "today." So we can construct a reduction of TB whose modeling language contains only context-insensitive claims. That reduction, model TB$^-$, is described in table 11.

L^- contains no sentences that go from certainty to less-than-certainty between t_0 and t_1, so we can use (LC) to derive diachronic verdicts for TB$^-$. We can also demonstrate that TB$^-$ is a proper reduction of TB (using $\mathsf{F} \in L^-$):

$$P_0(\mathit{Monday} \equiv \mathsf{F}) = 1 \qquad P_1(\mathit{Monday} \equiv \mathit{MonRed}) = 1$$

Table 10. Model TB

Story: Technicolor Beauty	ES: (1) $0 < P_0(Heads) < 1$
	(2) $P_0(\sim Monday) = 1$
T: Contains these times:	(3) $0 < P_1(\sim Monday) < 1$
t_0 Sunday night, after Beauty has heard the experiment described and made arrangements with her friend but before she is asleep.	(4) $0 < P_0(UpRed) < 1$
	(5) $P_1(UpRed) = 1$
	(6) $0 < P_0(Monday \equiv MonRed) < 1$
t_1 Monday morning, after Beauty awakens and sees the red paper.	(7) $P_1(Monday \equiv MonRed) = 1$
	(8) $P_0(Heads \supset [UpRed \equiv MonRed]) = 1$
L: Built on these atomic sentences, representing these claims:	(9) $P_1(Heads \supset [UpRed \equiv MonRed]) = 1$
	(10) $P_0(\sim Heads \supset UpRed) = 1$
Heads The coin comes up heads.	(11) $P_0(Heads \supset MonRed) < 1$
Monday Today is Monday.	(12) $P_0(Heads \supset \sim MonRed) < 1$
MonRed Monday is the red paper day.	(13) $P_1(Heads \supset MonRed) = 1$
UpRed Beauty awakens on the red paper day.	CGS: $\langle C_1 - C_0 \rangle \dashv\vdash UpRed \,\&\, (Monday \equiv MonRed) \,\&\, (Heads \supset MonRed)$

Table 11. Model TB$^-$

Story: Technicolor Beauty	*UpRed* Beauty awakens on the red paper day.
T^-: Contains these times:	
t_0 Sunday night, after Beauty has heard the experiment described and made arrangements with her friend but before she is asleep.	ES: (1) $0 < P_0^-(Heads) < 1$
	(2) $0 < P_0^-(UpRed) < 1$
	(3) $P_1^-(UpRed) = 1$
	(4) $P_0^-(Heads \supset [UpRed \equiv MonRed]) = 1$
t_1 Monday morning, after Beauty awakens and sees the red paper.	(5) $P_1^-(Heads \supset [UpRed \equiv MonRed]) = 1$
	(6) $P_0^-(\sim Heads \supset UpRed) = 1$
	(7) $P_0^-(Heads \supset MonRed) < 1$
L^-: Built on these atomic sentences, representing these claims:	(8) $P_0^-(Heads \supset \sim MonRed) < 1$
	(9) $P_1^-(Heads \supset MonRed) = 1$
Heads The coin comes up heads.	
MonRed Monday is the red paper day.	CGS: $\langle C_1 - C_0 \rangle \dashv\vdash UpRed \,\&\, (Heads \supset MonRed)$

So by (PEP) the diachronic verdicts of TB$^-$ will also be verdicts of TB. In particular, appendix B derives this diachronic verdict of TB:

$$P_1(\mathit{Heads}) = \frac{P_0(\mathit{MonRed} \mid \mathit{Heads}) \cdot P_0(\mathit{Heads})}{P_0(\mathit{MonRed} \mid \mathit{Heads}) \cdot P_0(\mathit{Heads}) + 1 - P_0(\mathit{Heads})}. \quad (14)$$

Since Monday is the red paper day just in case the die roll comes out odd, equation (14) expresses Beauty's Monday morning degree of belief in heads in terms of two values: her Sunday night degree of belief that the coin will come up heads, and her Sunday night degree of belief that the die roll will come out odd conditional on the coin's coming up heads. With a bit of algebra and the fact that both these degrees of belief are nonextreme (see appendix B), equation (14) yields

$$P_1(\mathit{Heads}) < P_0(\mathit{Heads}). \quad (15)$$

This analysis of the Sleeping Beauty Problem adds a feature (the colored papers) that gives Beauty a uniquely denoting context-insensitive expression on Monday morning for "today."[20] This allows us to relate Beauty's Sunday night degrees of belief to her Monday morning degrees of belief directly, without working through the intermediary of Monday night. At the same time, we have carefully kept the colored papers apparatus independent of the coin flip, making the requirements on Beauty's Monday morning degree of belief in heads in Technicolor Beauty the same as the requirements in the original Sleeping Beauty Problem.

Equation (15) therefore recovers our verdict from section 2.2 that ideal rationality requires Beauty's Monday morning degree of belief in heads to be less than her Sunday night degree of belief. This is sufficient to show that Lewis's and Bostrom's solutions to the Sleeping Beauty Problem are incorrect.

But the Technicolor Beauty analysis also yields a stronger result than we obtained from SB0 and SB1. Since Beauty is certain on Sunday night that the coin flip and the die roll are fair, independent chance events, the Principal Principle allows us to derive $P_0(\mathit{Heads}) = 1/2$ and $P_0(\mathit{MonRed} \mid \mathit{Heads}) = 1/2$. Substituting these values into equation

20. Kenny Easwaran suggested colored papers to me as a way of giving Beauty a uniquely denoting context-insensitive expression for "today." Brian Kierland and Bradley Monton (2005) also note that the Sleeping Beauty Problem does not require Beauty's awakenings to be subjectively indistinguishable, and they suggest a color-coding idea with pajamas similar to the colored papers apparatus here.

(14) yields

$$P_1(\mathit{Heads}) = \frac{1}{3}. \tag{16}$$

We now know the precise degree of belief ideal rationality requires Beauty to assign to heads on Monday morning. We have recovered Elga's answer to the Sleeping Beauty Problem without invoking an indifference principle and without applying the Principal Principle to Beauty's degrees of belief after Sunday night.

2.6. An Objection to This Solution

The strongest objection to our Technicolor Beauty analysis is that it isn't a solution to the Sleeping Beauty Problem at all. This objection grants that when Beauty awakens Monday morning and sees a red paper, ideal rationality requires her to assign a degree of belief of one-third to heads. However, the objection claims that because of the additional features of the Technicolor Beauty story, Beauty's required t_1 degree of belief in heads in Technicolor Beauty does not match her required t_1 degree of belief in heads in the original Sleeping Beauty Problem.

There are two times in Technicolor Beauty when Beauty gains beliefs she does not have in the original problem. The first time is when her friend agrees on Sunday night to place the colored papers. But this extra information about her friend's future behavior does not displace the original problem's requirements on Beauty's Sunday night degrees of belief concerning heads. So the focus of the objection must be on the second time: when Beauty awakens Monday morning and sees the red piece of paper. The concern is that information about which colored papers she gets to see alters the requirements on Beauty's Monday morning degree of belief in heads.

Let's suppose there is a small period of time after Beauty awakens Monday morning but before she sees the red piece of paper—call it $t_{0.5}$. If the objector grants that the extra Sunday night information in Technicolor Beauty does not disrupt the original problem's requirements on Beauty's degrees of belief, he should grant that Beauty's required degree of belief in heads at $t_{0.5}$ in Technicolor Beauty equals her required degree of belief in heads at t_1 in the original problem.

While I won't do so here, we could easily lay out a model TB* whose time set consists of $t_{0.5}$ and t_1 and whose modeling language L^* is identical to that of TB. In TB* we have $\langle C^*_1 - C^*_{0.5}\rangle \dashv\vdash \mathit{UpRed}$ &

(*Monday* ≡ *MonRed*) & (*Heads* ⊃ *MonRed*). Since Beauty is certain none of the claims represented in L^* changes truth-value between $t_{0.5}$ and t_1, none of the sentences in L^* goes from certainty to less-than-certainty between those two times. So we can apply (LC) to obtain

$$P_1^*(\mathit{Heads}) = P_{0.5}^*(\mathit{Heads} \mid \mathit{UpRed} \,\&\, [\mathit{Monday} \equiv \mathit{MonRed}] \,\&\, [\mathit{Heads} \supset \mathit{MonRed}]). \tag{17}$$

The objection we are considering grants that ideal rationality requires $P_1^*(\mathit{Heads}) = 1/3$. At $t_{0.5}$, Beauty is certain that *UpRed* & (*Monday* ≡ *MonRed*) & (*Heads* ⊃ *MonRed*) represents a claim with the same truth-value as "Today is the red paper day." So by substitution (see section 1.6), equation (17) tells us that ideal rationality requires Beauty to assign a $t_{0.5}$ degree of belief of one-third to heads conditional on the supposition that she will soon see a red piece of paper.

In section 2.5 we supposed without loss of generality that Beauty sees a *red* piece of paper on Monday. We could repeat the analysis of this section and the last for the blue-Monday version of Technicolor Beauty. Ideal rationality would still require Beauty to assign a degree of belief of one-third to heads at t_1 in that version, and we could derive an equation like equation (17) but with *MonRed* negated throughout. In the blue-Monday version ideal rationality would require Beauty to assign a $t_{0.5}$ degree of belief of one-third to heads conditional on the supposition that she will soon see a *blue* piece of paper.

Beauty's $t_{0.5}$ information in the blue-Monday version of Technicolor Beauty is identical to her $t_{0.5}$ information in the red-Monday version. Thus in the red-Monday version Beauty is required at $t_{0.5}$ to assign one-third to heads conditional on the supposition that she will soon see a blue piece of paper. But at $t_{0.5}$ Beauty is certain she will soon see either a red piece of paper or a blue piece of paper (but not both), so at $t_{0.5}$ she is required to assign one-third to heads conditional both on the supposition that she will soon see a red piece of paper and on the supposition that she will not. By our synchronic constraints, this entails that ideal rationality requires Beauty to assign an *unconditional* degree of belief of one-third to heads at $t_{0.5}$.

The objector has already granted that Beauty's required $t_{0.5}$ degree of belief in heads equals the t_1 degree of belief in heads required of her in the original Sleeping Beauty Problem. Thus he must now concede that ideal rationality requires Beauty to assign a degree of belief of one-third to heads at t_1 in the original problem. Beauty's extra

information about colored papers in the Technicolor Beauty story does not alter the original problem's requirements on her t_1 degrees of belief.

2.7. The HT Approach

The current formal epistemology literature is rife with analyses of the Sleeping Beauty Problem. The tendency is to lay out the original problem—or a problem asserted to be analogous to the original—then apply a method that feels intuitively reasonable for that case. There is rarely any discussion of how the method might generalize or of results it yields for other examples.

A welcome exception to this tendency is Joseph Halpern's (2005) presentation of an updating policy he calls the "HT approach."[21] Roughly, the HT approach directs an agent to update first by assigning a new credence distribution over uncentered worlds that is the conditionalization of her old uncentered world distribution on what she has learned, then second by distributing her credence in each uncentered world among the centered worlds associated with it.[22] Applied to the Sleeping Beauty Problem, the HT approach requires Beauty to assign one-half to heads on Monday morning and one-half to heads on Monday night (after she has learned it is Monday). These assignments are consistent with the Relevance-Limiting Thesis—in fact, the HT approach yields the Relevance-Limiting Thesis as a theorem. The HT approach also yields a different answer for the Sleeping Beauty Problem than it does for Technicolor Beauty; applied to the latter, it requires Beauty to assign one-third to heads on Monday morning. We criticized these positions in sections 2.3 and 2.6, but the criticisms there were based on our modeling framework and so do not apply to someone like Halpern who can reject our framework in favor of another.

Nevertheless, the gap between its answers to the Sleeping Beauty Problem and to Technicolor Beauty is a serious strike against the HT approach. The colored papers in Technicolor Beauty create a probabilistically simple way for Beauty to distinguish one awakening from the other. But Beauty's awakenings could become subjectively distinguishable in much subtler ways. If she follows the HT approach,

21. The HT approach extends a formal model developed by Halpern and Mark Tuttle.

22. Meacham (2008) presents an alternative framework that yields verdicts similar to those of the HT approach. Meacham's framework shares the features of the HT approach I discuss here.

Beauty will awaken Monday morning with a one-half degree of belief in heads. But the moment *anything* occurs that she is less-than-certain will also occur on her other awakening, her degree of belief in heads dips below one-half. The HT approach keeps Beauty's self-locating beliefs irrelevant at the cost of making the most trivial details—a fly that buzzes into the room, a cough in the middle of one of the experimenter's questions—relevant to her degree of belief in heads.

There is also a more fundamental difference between our framework and the HT approach. Distinguishing centered from uncentered worlds was an important step in the philosophy of language. But that is insufficient reason to build the distinction into the basic structure of a formal modeling system in epistemology. Our methodology has been to start with examples (The Die and Sleeping In) in which the requirements of ideal rationality are uncontroversial. We developed a generally applicable, precisely stated modeling framework whose verdicts match those requirements. In the process, we found that self-locating claims can generate exceptions to the traditional Conditionalization updating rule when they result in a loss of certainty. However (as we saw with the sentence *Monday* in section 2.1), when self-locating claims do not go from certainty to less-than-certainty between two times, they can be conditionalized upon just like non-self-locating claims. At the same time (as we saw in section 2.3), a non-self-locating claim can require special handling in a model when it goes from certainty to less-than-certainty due to memory loss.

The important distinction for epistemic modeling is not between self-locating and non-self-locating claims, but between claims that go from certainty to less-than-certainty and those that do not. By explicitly building a difference between self-locating and non-self-locating beliefs into the framework and then offering a distinct updating rule for each, the HT approach overreacts to an epistemic complication that only sometimes results from self-locating claims. Our modeling framework has no need to distinguish sentences that represent self-locating claims (or centered worlds) from those that do not, because the critical feature (loss of certainty) can be represented in the framework syntactically. Thus while some of our arguments for the framework have involved intuitions about semantics, the formal principles of the framework itself are purely syntactical. This makes our framework extremely general; for example, we might try applying it to stories involving context-sensitive claims that are not self-locating.

Having built our framework using uncontroversial examples, we can apply it to controversial stories like the Sleeping Beauty Problem. When we do, we find that ideal rationality requires Beauty to decrease her degree of belief in heads between Sunday night and Monday morning. Because she was certain in advance that she would be awakening in precisely the conditions she finds on Monday morning, Beauty learns only self-locating claims between those two times. Yet ideal rationality requires her to alter her degree of belief in the non-self-locating claim that the coin comes up heads. So the Sleeping Beauty Problem is a counterexample to the Relevance-Limiting Thesis: it *can* be rational for an agent who learns only self-locating information to respond by altering a non-self-locating degree of belief.

Ideally rational degrees of belief can stand in subtle and complex relevance relations. In modeling these relations, we should not start by stipulating high barriers to inference between claims of various kinds (the Relevance-Limiting Thesis), nor should we assume there are different updating rules for self-locating and non-self-locating claims (the HT approach). Instead, we should develop a modeling framework that is general enough to model a wide variety of stories, sufficiently formal that there is no debate about what it says in any given case, and responsive to our intuitions about settled and obvious examples. We can then let our models teach *us* about the relevance relations.

Appendix A

Call M^+ a **perfect expansion** of M just in case M^+ is an expansion of M and

$$(\forall y \in L^+)(\exists x \in L)(\forall t_k \in T^+)(P_k^+(x \equiv y) = 1).$$

This definition differs from that of a proper expansion in the order of the last two quantifiers. Every perfect expansion is a proper expansion, but the converse does not hold.

In this appendix we will work within a modeling framework that is identical to the one presented in the body of the essay except that it lacks (PEP) as a systematic constraint—we'll call it the **non-(PEP) framework**. Our goal will be to prove in the non-(PEP) framework that if M^+ is a perfect expansion of M, the analogue for M^+ of any verdict of M is a verdict of M^+.

Any verdict of M can be derived algebraically from a set of premises each of which is either an extrasystematic constraint on M or

an instance in M of one of the non-(PEP) framework's systematic constraints. If the analogue for M^+ of every such premise is a verdict of M^+, every verdict of M will have an analogue for M^+ that is a verdict of M^+. By the definition of an expansion, if M^+ is an expansion of M every extrasystematic constraint on M has an analogue for M^+ that is an extrasystematic constraint on M^+ and therefore a verdict of M^+. So we focus our attention on analogues of instances of systematic constraints (1) through (5).

LEMMA A.1
If $P_k(x) = 1$ is a verdict of model M, then $P_k(x) = 1$ is an extrasystematic constraint on model M.

Proof
Suppose $P_k(x) = 1$ is a verdict of M. There is either an extrasystematic constraint on M that $P_k(x) = 1$ or an extrasystematic constraint on M that $P_k(x) < 1$. Suppose for *reductio* that the latter is true. Then M has contradictory verdicts. Systematic constraints (1) through (5) cannot derive contradictory verdicts from a consistent set of extrasystematic constraints, so M's extrasystematic constraints must be inconsistent. But we assume that the stipulations of any story we analyze will be consistent. So we have a contradiction, and there must be an extrasystematic constraint on M that $P_k(x) = 1$. □

LEMMA A.2
If $P_k(x) < 1$ is a verdict of model M, then $P_k(x) < 1$ is an extrasystematic constraint on model M.

Proof
Parallel to proof of lemma (A.1), but with a *reductio* of the first disjunct in the second sentence. □

THEOREM A.3
If M^+ is an expansion of M, any instance of systematic constraints (1) through (4) in M has an analogue for M^+ that is a verdict of M^+.

Proof
Systematic constraint (1) can be read

$$(\forall t_k \in T)(\forall x \in L)(P_k(x) \geq 0).$$

By the definition of an expansion, $T = T^+$ and $L \subseteq L^+$. So any instance of systematic constraint (1) in M will have an analogue for M^+ that is an instance of systematic constraint (1) in M^+ and therefore a verdict of M^+.

Systematic constraint (2) can be read

$$(\forall t_k \in T)(\forall x \in L)([x \text{ is a tautology}] \supset P_k(x) = 1).$$

If x is a tautology in L, it will also be a tautology in L^+, so a similar argument applies.

Systematic constraint (3) can be read

$$(\forall t_k \in T)(\forall x \in L)(\forall y \in L)([x, y \text{ are mutually exclusive}] \supset P_k(x \vee y) = P_k(x) + P_k(y)).$$

If x and y are mutually exclusive in L, they will also be mutually exclusive in L^+, so a similar argument applies.

The first sentence of systematic constraint (4) can be read

$$(\forall t_k \in T)(\forall x \in L)(\forall y \in L)(P_k(y) > 0 \supset P_k(x \mid y) = \tfrac{P_k(x\&y)}{P_k(y)}).$$

Suppose $P_k(y) > 0$ is a verdict of M. Then $P_k(\sim y) < 1$ is a verdict of M, and by lemma (A.2), $P_k(\sim y) < 1$ is an extrasystematic constraint on M. Since M^+ is an expansion of M, $P_k^+(\sim y) < 1$ also is an extrasystematic constraint on M^+, so $P_k^+(y) > 0$ is a verdict of M^+. Thus if the antecedent of the conditional in systematic constraint (4) is met in M, its analogue will be met in M^+, and the instance of the consequent in M will have an analogue for M^+ that is a verdict of M^+.

If $P_k(y) = 0$ is a verdict of M, $P_k(\sim y) = 1$ also will be a verdict of M, by lemma (A.1), $P_k(\sim y) = 1$ will be an extrasystematic constraint on M, and $P_k^+(\sim y) = 1$ will be an extrasystematic constraint on M^+. So if "$P_k(x \mid y)$ is undefined" is a verdict of M, then "$P_k^+(x \mid y)$ is undefined" will be a verdict of M^+. □

COROLLARY A.4

If M^+ is a proper expansion of M, any verdict of M that can be derived from just extrasystematic constraints and synchronic systematic constraints has an analogue for M^+ that is a verdict of M^+.

LEMMA A.5
If M^+ is a perfect expansion of M and $C_j \subseteq C_k$ for some $t_j, t_k \in T$, then $C_j^+ \subseteq C_k^+$.

Proof
Suppose M^+ is a perfect expansion of M and $C_j \subseteq C_k$ for some $t_j, t_k \in T$, and suppose for *reductio* that $C_j^+ \nsubseteq C_k^+$. Then there exists $y \in L^+$ such that $P_j^+(y) = 1$ and $P_k^+(y) < 1$. Since M^+ is a perfect expansion of M, there exists $x \in L$ such that for any t_i, $P_i^+(x \equiv y) = 1$. By substitution (see section 1.6), $P_j^+(x) = 1$ and $P_k^+(x) < 1$. By lemmas (A.1) and (A.2), each of these is an extrasystematic constraint on M^+, so their analogues are also extrasystematic constraints on M. But then $x \in C_j$ and $x \notin C_k$, so $C_j \nsubseteq C_k$, and we have a contradiction. □

THEOREM A.6
If M^+ is a perfect expansion of M, any instance of systematic constraint (5) in M has an analogue for M^+ that is a verdict of M^+.

Proof
Suppose M^+ is a perfect expansion of M. An instance of systematic constraint (5) in M takes the form

$$P_k(z) = P_j(z \mid \langle C_k - C_j \rangle), \tag{18}$$

where $C_j \subseteq C_k$.

By lemma (A.5), $C_j^+ \subseteq C_k^+$. So systematic constraint (5) yields the verdict

$$P_k^+(z) = P_j^+(z \mid \langle C_k^+ - C_j^+ \rangle). \tag{19}$$

Define the set $S = (C_k^+ - C_j^+) \cap L$, and consider an arbitrary $y \in C_k^+ - C_j^+$. For that y, $P_j^+(y) < 1$ and $P_k^+(y) = 1$. Since M^+ is a perfect expansion of M, there exists $x \in L$ such that $P_j^+(x \equiv y) = 1$ and $P_k^+(x \equiv y) = 1$. By substitution, $P_j^+(x) < 1$ and $P_k^+(x) = 1$. So $x \in S$. Since y was arbitrarily selected from $C_k^+ - C_j^+$, we have now shown that for every $y \in C_k^+ - C_j^+$ there exists an $x \in S$ such that $P_j^+(x \equiv y) = 1$. By our synchronic constraints, we have

$$P_j^+(\langle C_k^+ - C_j^+ \rangle \equiv \langle S \rangle) = 1. \tag{20}$$

Applying substitution to equations (19) and (20) yields

$$P_k^+(z) = P_j^+(z \mid \langle S \rangle). \tag{21}$$

Suppose a sentence x is in S. Then $x \in L$, $P_j^+(x) < 1$, and $P_k^+(x) = 1$. By lemmas (A.2) and (A.1), the latter must both be extrasystematic constraints on M^+. So their analogues for M are extrasystematic constraints on M, and $x \in C_k - C_j$. Moving in the other direction, suppose x is in $C_k - C_j$. Then $x \in L$, $P_j(x) < 1$, and $P_k(x) = 1$. Both of the latter are extrasystematic constraints on M, and their analogues are extrasystematic constraints on M^+. So $x \in S$. Thus $S = C_k - C_j$, and equation (21) is the analogue for M^+ of equation (18). We have shown that any instance of systematic constraint (5) in M has an analogue for M^+ that is a verdict of M^+. □

COROLLARY A.7
If M^+ is a proper expansion of M and no sentences go from a credence of 1 *to a credence less than* 1 *in M^+, then any verdict of M has an analogue for M^+ that is a verdict of M^+.*

Proof
Let t_0 be the earliest time in T^+. Since M^+ is a proper expansion of M, for any $y \in L^+$ there exists an $x \in L$ such that $P_0^+(x \equiv y) = 1$. Since no sentences go from a credence of 1 to a credence less than 1 in M^+, for any $t_k \in T^+$ we have $P_k^+(x \equiv y) = 1$. Thus M^+ is a perfect expansion of M, and previous results in this appendix apply. □

Appendix B

First, an explanation of some extrasystematic constraints on model TB (section 2.5, table 10): (6) comes from the fact that Beauty is certain on Sunday that it is not Monday, but uncertain whether Monday will be the red paper day. At t_1 Beauty is certain that it is the red paper day, so she is also certain that it is Monday just in case Monday is the red paper day, yielding (7). (8), (9), and (10) stem from Beauty's certainty that if the coin comes up heads she awakens only on Monday, while if it comes up tails she awakens both days. Neither *Heads* $\supset$ *MonRed* nor *Heads* $\supset \sim$*MonRed* is stipulated as certain at t_0 by the story or entailed by claims stipulated as certain; this accounts for (11) and (12). However, *Heads* $\supset$ *MonRed* is entailed by *UpRed* and *Heads* $\supset$ (*UpRed* $\equiv$ *MonRed*), both of which are certain at t_1; this yields (13).

While TB has sentences going from certainty to less-than-certainty, model TB^- (table 11) does not, so (LC) yields

$$P_1^-(Heads) = P_0^-(Heads \mid UpRed \,\&\, [Heads \supset MonRed]). \tag{22}$$

As the main text points out, TB is a proper expansion of TB^-, so (PEP) yields

$$P_1(Heads) = P_0(Heads \mid UpRed \,\&\, [Heads \supset MonRed]). \tag{23}$$

At t_0 Beauty is certain that $UpRed \,\&\, (Heads \supset MonRed)$ has the same truth-value as *UpRed*. (This follows from the certainty described in extrasystematic constraint (8) on TB.) By substitution (see section 1.6), equation (23) becomes

$$P_1(Heads) = P_0(Heads \mid UpRed). \tag{24}$$

(Intuitively, the crucial piece of information Beauty learns between Sunday night and Monday morning is that she gets to awaken on the red paper day; on Sunday night she wasn't certain she would see the red paper.)

Applying Bayes's Theorem (which follows from our synchronic constraints) to equation (24) yields

$$P_1(Heads) = \frac{P_0(UpRed \mid Heads) \cdot P_0(Heads)}{P_0(UpRed \mid Heads) \cdot P_0(Heads) + P_0(UpRed \mid \sim Heads) \cdot P_0(\sim Heads)}. \tag{25}$$

Extrasystematic constraint (8) yields $P_0(UpRed \mid Heads) = P_0(MonRed \mid Heads)$, while extrasystematic constraint (10) yields $P_0(UpRed \mid \sim Heads) = 1$. So we have

$$P_1(Heads) = \frac{P_0(MonRed \mid Heads) \cdot P_0(Heads)}{P_0(MonRed \mid Heads) \cdot P_0(Heads) + 1 - P_0(Heads)}. \tag{26}$$

Extrasystematic constraint (1) guarantees that $P_0(Heads)$ is nonextreme, while extrasystematic constraints (11) and (12) guarantee the same for $P_0(MonRed \mid Heads)$. With a bit of algebra, equation (26) then yields

$$P_1(Heads) < P_0(Heads). \tag{27}$$

References

Arntzenius, F. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy* 100: 356–70.

Bostrom, N. 2007. "Sleeping Beauty and Self-Location: A Hybrid Model." *Synthese* 157: 59–78.

Elga, A. 2000. "Self-Locating Belief and the Sleeping Beauty Problem." *Analysis* 60: 143–47.

———. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research* 69: 383–96.

Garber, D. 1983. "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory." In *Testing Scientific Theories*, ed. J. Earman, 99–132. Minneapolis: University of Minnesota Press.

Hájek, A. 2003. "What Conditional Probability Could Not Be." *Synthese* 137: 273–323.

Halpern, J. Y. 2005. "Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems." In *Oxford Studies in Epistemology*, vol. 1, ed. T. Gendler and J. Hawthorne, 111–42. Oxford: Oxford University Press.

Jeffrey, R. C. 1983. *The Logic of Decision*, 2nd ed. Chicago: University of Chicago Press.

Kierland, B., and B. Monton. 2005. "Minimizing Inaccuracy for Self-Locating Beliefs." *Philosophy and Phenomenological Research* 70: 384–95.

Lance, M. N. 1995. "Subjective Probability and Acceptance." *Philosophical Studies* 77: 147–79.

Lewis, D. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review* 88: 513–43.

———. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, vol. 2, ed. R. C. Jeffrey, 263–94. Berkeley: University of California Press.

———. 2001. "Sleeping Beauty: Reply to Elga." *Analysis* 61: 171–76.

Meacham, C. J. G. 2008. "Sleeping Beauty and the Dynamics of De Se Beliefs." *Philosophical Studies* 138: 245–70.

Schervish, M. J., T. Seidenfeld, and J. Kadane. 2004. "Stopping to Reflect." *Journal of Philosophy* 101: 315–22.

Talbott, W. J. 1991. "Two Principles of Bayesian Epistemology." *Philosophical Studies* 62: 135–50.

Titelbaum, M. G. 2008. "Quitting Certainties: A Doxastic Modeling Framework." PhD diss., University of California, Berkley.

van Fraassen, B. C. 1995. "Belief and the Problem of Ulysses and the Sirens." *Philosophical Studies* 77: 7–37.